Computational Techniques for the Analysis of Small Signals in High-Statistics Neutrino Oscillation Experiments

M. G. Aartsen[q], M. Ackermann[bg], J. Adams[q], J. A. Aguilar[m], M. Ahlers[u], M. Ahrens[ax], I. Al Samarai[z], D. Altmann[y], K. Andeen[ak], T. Anderson[bd], I. Ansseau[m], G. Anton[y], C. Argüelles[o], T. C. Arlen[bd], J. Auffenberg[a], S. Axani[o], H. Bagherpour[q], X. Bai[au], A. Balagopal V.[ac], J. P. Barron[x], I. Bartos[aq], S. W. Barwick[ab], V. Baum[ai], R. Bay[h], J. J. Beatty[s,t], J. Becker Tjus[k], K.-H. Becker[bf], S. BenZvi[aw], D. Berley[r], E. Bernardini[bg], D. Z. Besson[ad], G. Binder[i,h], D. Bindig[bf], E. Blaufuss[r], S. Blot[bg], C. Bohm[ax], M. Bohmer[al], M. Börner[v], F. Bos[k], S. Böser[ai], O. Botner[be], E. Bourbeau[u], J. Bourbeau[ah], F. Bradascio[bg], J. Braun[ah], M. Brenzke[a], H.-P. Bretz[bg], S. Bron[z], J. Brostean-Kaiser[bg], A. Burgman[be], R. S. Busse[ah], T. Carver[z], E. Cheung[r], D. Chirkin[ah], A. Christov[z], K. Clark[ae], L. Classen[an], G. H. Collin[o], J. M. Conrad[o], P. Coppin[n], P. Correa[n], D. F. Cowen[bd,bc], R. Cross[aw], P. Dave[f], M. Day[ah], J. P. A. M. de André[w], C. De Clercq[n], J. J. DeLaunay[bd], H. Dembinski[ao], S. De Ridder[aa], P. Desiati[ah], K. D. de Vries[n], G. de Wasseige[n], M. de With[j], T. DeYoung[w], J. C. Díaz-Vélez[ah], V. di Lorenzo[ai], H. Dujmovic[az], J. P. Dumm[ax], M. Dunkman[bd], M. A. DuVernois[ah], E. Dvorak[au], B. Eberhardt[ai], T. Ehrhardt[ai], B. Eichmann[k], P. Eller[bd], R. Engel[ac], J. J. Evans[aj], P. A. Evenson[ao], S. Fahey[ah], A. R. Fazely[g], J. Felde[r], K. Filimonov[h], C. Finley[ax], S. Flis[ax], A. Franckowiak[bg], E. Friedman[r], A. Fritz[ai], T. K. Gaisser[ao], J. Gallagher[ag], A. Gartner[al], L. Gerhardt[i], R. Gernhaeuser[al], K. Ghorbani[ah], W. Giang[x], T. Glauch[al], T. Glüsenkamp[y], A. Goldschmidt[i], J. G. Gonzalez[ao], D. Grant[x], Z. Griffith[ah], C. Haack[a], A. Hallgren[be], F. Halzen[ah], K. Hanson[ah], J. Haugen[ah], A. Haungs[ac], D. Hebecker[j], D. Heereman[m], K. Helbing[bf], R. Hellauer[r], F. Henningsen[al], S. Hickford[bf], M. Hieronymus[ai], J. Hignight[w], G. C. Hill[b], K. D. Hoffman[r], B. Hoffmann[ac], R. Hoffmann[bf], T. Hoinka[v], B. Hokanson-Fasig[ah], K. Holzapfel[al], K. Hoshina[ah,ba], F. Huang[bd], M. Huber[al], T. Huber[ac], T. Huege[ac], K. Hultqvist[ax], M. Hünnefeld[v], R. Hussain[ah], S. In[az], N. Iovine[m], A. Ishihara[p], E. Jacobi[bg], G. S. Japaridze[e], M. Jeong[az], K. Jero[ah], B. J. P. Jones[d], P. Kalaczynski[a], O. Kalekin[y], W. Kang[az], D. Kang[ac], A. Kappes[an], D. Kappesser[ai], T. Karg[bg], A. Karle[ah], T. Katori[af], U. Katz[y], M. Kauer[ah], A. Keivani[bd], J. L. Kelley[ah], A. Kheirandish[ah], J. Kim[az], M. Kim[p], T. Kintscher[bg], J. Kiryluk[ay], T. Kittler[y], S. R. Klein[i,h], R. Koirala[ao], H. Kolanoski[j], L. Köpke[ai], C. Kopper[x], S. Kopper[bb], J. P. Koschinsky[a], D. J. Koskinen[u], M. Kowalski[j,bg], C. B. Krauss[x], K. Krings[al], M. Kroll[k], G. Krückl[ai], S. Kunwar[bg], N. Kurahashi[at], T. Kuwabara[p], A. Kyriacou[b], M. Labare[aa], J. L. Lanfranchi[bd], M. J. Larson[u], F. Lauber[bf], D. Lennarz[w], K. Leonard[ah], M. Lesiak-Bzdak[ay], A. Leszczynska[ac], M. Leuermann[a], Q. R. Liu[ah], E. Lohfink[ai], J. LoSecco[ar], C. J. Lozano Mariscal[an], L. Lu[p], J. Lünemann[n], W. Luszczak[ah], J. Madsen[av], G. Maggi[n], K. B. M. Mahn[w], S. Mancina[ah], S. Mandalia[af], S. Marka[aq], Z. Marka[aq], R. Maruyama[ap], K. Mase[p], R. Maunu[r], K. Meagher[m], M. Medici[u], M. Meier[v], T. Menne[v], G. Merino[ah], T. Meures[m], S. Miarecki[i,h], J. Micallef[w], G. Momenté[ai], T. Montaruli[z], R. W. Moore[x], M. Moulai[o], R. Nahnhauer[bg], P. Nakarmi[bb], U. Naumann[bf], G. Neer[w], H. Niederhausen[ay], S. C. Nowicki[x], D. R. Nygren[i], A. Obertacke Pollmann[bf], M. Oehler[ac], A. Olivas[r],

A. O'Murchadha[m], E. O'Sullivan[ax], A. Palazzo[am], T. Palczewski[i,h], H. Pandya[ao], D. V. Pankova[bd], L. Papp[al], P. Peiffer[ai], J. A. Pepper[bb], C. Pérez de los Heros[be], T. C. Petersen[u], D. Pieloth[v], E. Pinat[m], J. L. Pinfold[x], M. Plum[ak], P. B. Price[h], G. T. Przybylski[i], C. Raab[m], L. Rädel[a], M. Rameez[u], L. Rauch[bg], K. Rawlins[c], I. C. Rea[al], R. Reimann[a], B. Relethford[at], M. Relich[p], M. Renschler[ac], E. Resconi[al], W. Rhode[v], M. Richman[at], M. Riegel[ac], S. Robertson[b], M. Rongen[a], C. Rott[az], T. Ruhe[v], D. Ryckbosch[aa], D. Rysewyk[w], I. Safa[ah], T. Sälzer[a], S. E. Sanchez Herrera[x], A. Sandrock[v], J. Sandroos[ai], P. Sandstrom[ah], M. Santander[bb], S. Sarkar[u,as], S. Sarkar[x], K. Satalecka[bg], H. Schieler[ac], P. Schlunder[v], T. Schmidt[r], A. Schneider[ah], S. Schoenen[a], S. Schöneberg[k], F. G. Schröder[ac], L. Schulte[l], L. Schumacher[a], S. Sclafani[at], D. Seckel[ao], S. Seunarine[av], M. H. Shaevitz[aq], J. Soedingrekso[v], D. Soldin[ao], S. Söldner-Rembold[aj], M. Song[r], G. M. Spiczak[av], C. Spiering[bg], J. Stachurska[bg], M. Stamatikos[s], T. Stanev[ao], A. Stasik[bg], R. Stein[bg], J. Stettner[a], A. Steuer[ai], T. Stezelberger[i], R. G. Stokstad[i], A. Stößl[p], N. L. Strotjohann[bg], T. Stuttard[u], G. W. Sullivan[r], M. Sutherland[s], I. Taboada[f], A. Taketa[ba], H. K. M. Tanaka[ba], J. Tatar[i,h], F. Tenholt[k], S. Ter-Antonyan[g], A. Terliuk[bg], S. Tilav[ao], P. A. Toale[bb], M. N. Tobin[ah], C. Tönnis[az], S. Toscano[n], D. Tosi[ah], M. Tselengidou[y], C. F. Tung[f], A. Turcati[al], C. F. Turley[bd], B. Ty[ah], E. Unger[be], M. Usner[bg], J. Vandenbroucke[ah], W. Van Driessche[aa], D. van Eijk[ah], N. van Eijndhoven[n], S. Vanheule[aa], J. van Santen[bg], D. Veberic[ac], E. Vogel[a], M. Vraeghe[aa], C. Walck[ax], A. Wallace[b], M. Wallraff[a], F. D. Wandler[x], N. Wandkowsky[ah], A. Waza[a], C. Weaver[x], A. Weindl[ac], M. J. Weiss[bd], C. Wendt[ah], J. Werthebach[ah], S. Westerhoff[ah], B. J. Whelan[b], K. Wiebe[ai], C. H. Wiebusch[a], L. Wille[ah], D. R. Williams[bb], L. Wills[at], M. Wolf[ah], J. Wood[ah], T. R. Wood[x], E. Woolsey[x], K. Woschnagg[h], G. Wrede[y], S. Wren[aj], D. L. Xu[ah], X. W. Xu[g], Y. Xu[ay], J. P. Yanez[x], G. Yodh[ab], S. Yoshida[p], T. Yuan[ah]

[a] *III. Physikalisches Institut, RWTH Aachen University, D-52056 Aachen, Germany*
[b] *Department of Physics, University of Adelaide, Adelaide, 5005, Australia*
[c] *Dept. of Physics and Astronomy, University of Alaska Anchorage, 3211 Providence Dr., Anchorage, AK 99508, USA*
[d] *Dept. of Physics, University of Texas at Arlington, 502 Yates St., Science Hall Rm 108, Box 19059, Arlington, TX 76019, USA*
[e] *CTSPS, Clark-Atlanta University, Atlanta, GA 30314, USA*
[f] *School of Physics and Center for Relativistic Astrophysics, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[g] *Dept. of Physics, Southern University, Baton Rouge, LA 70813, USA*
[h] *Dept. of Physics, University of California, Berkeley, CA 94720, USA*
[i] *Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[j] *Institut für Physik, Humboldt-Universität zu Berlin, D-12489 Berlin, Germany*
[k] *Fakultät für Physik & Astronomie, Ruhr-Universität Bochum, D-44780 Bochum, Germany*
[l] *Physikalisches Institut, Universität Bonn, Nussallee 12, D-53115 Bonn, Germany*
[m] *Université Libre de Bruxelles, Science Faculty CP230, B-1050 Brussels, Belgium*
[n] *Vrije Universiteit Brussel (VUB), Dienst ELEM, B-1050 Brussels, Belgium*
[o] *Dept. of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[p] *Dept. of Physics and Institute for Global Prominent Research, Chiba University, Chiba 263-8522, Japan*
[q] *Dept. of Physics and Astronomy, University of Canterbury, Private Bag 4800, Christchurch, New Zealand*
[r] *Dept. of Physics, University of Maryland, College Park, MD 20742, USA*

[s]Dept. of Physics and Center for Cosmology and Astro-Particle Physics, Ohio State University, Columbus, OH 43210, USA
[t]Dept. of Astronomy, Ohio State University, Columbus, OH 43210, USA
[u]Niels Bohr Institute, University of Copenhagen, DK-2100 Copenhagen, Denmark
[v]Dept. of Physics, TU Dortmund University, D-44221 Dortmund, Germany
[w]Dept. of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA
[x]Dept. of Physics, University of Alberta, Edmonton, Alberta, Canada T6G 2E1
[y]Erlangen Centre for Astroparticle Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, D-91058 Erlangen, Germany
[z]Département de physique nucléaire et corpusculaire, Université de Genève, CH-1211 Genève, Switzerland
[aa]Dept. of Physics and Astronomy, University of Gent, B-9000 Gent, Belgium
[ab]Dept. of Physics and Astronomy, University of California, Irvine, CA 92697, USA
[ac]Institut für Kernphysik, Karlsruhe Institute of Technology, D-76021 Karlsruhe, Germany
[ad]Dept. of Physics and Astronomy, University of Kansas, Lawrence, KS 66045, USA
[ae]SNOLAB, 1039 Regional Road 24, Creighton Mine 9, Lively, ON, Canada P3Y 1N2
[af]School of Physics and Astronomy, Queen Mary University of London, London E1 4NS, United Kingdom
[ag]Dept. of Astronomy, University of Wisconsin, Madison, WI 53706, USA
[ah]Dept. of Physics and Wisconsin IceCube Particle Astrophysics Center, University of Wisconsin, Madison, WI 53706, USA
[ai]Institute of Physics, University of Mainz, Staudinger Weg 7, D-55099 Mainz, Germany
[aj]School of Physics and Astronomy, The University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom
[ak]Department of Physics, Marquette University, Milwaukee, WI, 53201, USA
[al]Physik-department, Technische Universität München, D-85748 Garching, Germany
[am]Max-Planck-Institut für Physik (Werner Heisenberg Institut), Föhringer Ring 6, D-80805 München, Germany
[an]Institut für Kernphysik, Westfälische Wilhelms-Universität Münster, D-48149 Münster, Germany
[ao]Bartol Research Institute and Dept. of Physics and Astronomy, University of Delaware, Newark, DE 19716, USA
[ap]Dept. of Physics, Yale University, New Haven, CT 06520, USA
[aq]Columbia Astrophysics and Nevis Laboratories, Columbia University, New York, NY 10027, USA
[ar]Dept. of Physics, University of Notre Dame du Lac, 225 Nieuwland Science Hall, Notre Dame, IN 46556-5670, USA
[as]Dept. of Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK
[at]Dept. of Physics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
[au]Physics Department, South Dakota School of Mines and Technology, Rapid City, SD 57701, USA
[av]Dept. of Physics, University of Wisconsin, River Falls, WI 54022, USA
[aw]Dept. of Physics and Astronomy, University of Rochester, Rochester, NY 14627, USA
[ax]Oskar Klein Centre and Dept. of Physics, Stockholm University, SE-10691 Stockholm, Sweden
[ay]Dept. of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794-3800, USA
[az]Dept. of Physics, Sungkyunkwan University, Suwon 440-746, Korea
[ba]Earthquake Research Institute, University of Tokyo, Bunkyo, Tokyo 113-0032, Japan
[bb]Dept. of Physics and Astronomy, University of Alabama, Tuscaloosa, AL 35487, USA
[bc]Dept. of Astronomy and Astrophysics, Pennsylvania State University, University Park, PA 16802, USA
[bd]Dept. of Physics, Pennsylvania State University, University Park, PA 16802, USA
[be]Dept. of Physics and Astronomy, Uppsala University, Box 516, S-75120 Uppsala, Sweden
[bf]Dept. of Physics, University of Wuppertal, D-42119 Wuppertal, Germany
[bg]DESY, D-15738 Zeuthen, Germany

Abstract

The current and upcoming generation of Very Large Volume Neutrino Telescopes—collecting unprecedented quantities of neutrino events—can be used to explore subtle effects in oscillation physics, such as (but not restricted to) the neutrino mass ordering. The sensitivity of an experiment to these effects can be estimated from Monte Carlo simulations. With the very high number of events that will be collected, there is a trade-off between the computational expense of running such simulations and the inherent statistical uncertainty in the determined values. In such a scenario, it becomes impractical to produce and use adequately-sized sets of simulated events to use with traditional methods, such as Monte Carlo weighting. In this work we present a staged approach to the generation of binned event distributions in order to overcome these challenges. By combining multiple integration and smoothing techniques which address limited statistics from simulation it arrives at reliable analysis results using modest computational resources.

Keywords: Data Analysis, Monte Carlo, MC, Statistics, Smoothing, KDE, Neutrino, Neutrino Mass Ordering, Detector, VLVνT

1. Introduction

By virtue of their multi-megaton effective mass paired with the magnitude of the atmospheric neutrino flux, the next generation of Very Large Volume Neutrino Telescopes (VLVνTs) dedicated to neutrino oscillation physics, such as PINGU and ORCA [1, 2, 3], will record tens of thousands of GeV-scale neutrino interactions. These large-scale water or ice Cherenkov detectors do not have the ability to unambiguously distinguish between neutrino flavors and interaction types on an event-by-event basis. Even so, their high statistics data samples can be used to explore small effects such as the tau neutrino appearance rate, the ordering of the neutrino mass eigenstates (NMO), or potential neutrino physics beyond the Standard Model.

All such physics analyses are carried out by comparing the observed event distributions with predictions (hereafter referred to as *templates*) obtained from Monte Carlo (MC) simulations. The physical phenomena listed above will appear in the templates as deviations in event count as small as a few percent. An inherent problem when trying to quantify these deviations in high-statistics data sets is that the templates must be described with an accuracy better than the magnitude of the effect being investigated. A limiting factor to the accuracy is the amount of MC simulation available, which is in turn constrained by the availability of computing resources. This particularly applies during the design optimization phase of a planned experiment, which entails performance assessments of multiple detector variants.

Once an accurate template has been produced, extracting the relevant physical and systematic parameters typically proceeds via maximizing the likelihood of obtaining the

*analysis@icecube.wisc.edu

observed data under a given hypothesis. A common feature to all statistical methods is that the templates need to be generated for a multitude of parameter combinations, often thousands or even millions. This process needs to be accurate, but also fast, which typically prohibits the reproduction of the full MC sample for each template.

In this paper, we present an approach that allows us to obtain fast and accurate templates even from MC sets that are several orders of magnitude smaller than those necessary when using simpler methods. These methods and tools were used to calculate the expected sensitivities for atmospheric neutrino oscillation analyses with the proposed low-energy extension of the IceCube experiment [1, 2]. Throughout this paper, we will use the NMO analysis for a generic VLVνT as an example to illustrate our methods, though it is applicable in a wider context. Section 2 details the computational challenge at hand. Our approach to overcome this challenge is presented in Section 3 and Section 4, followed by a discussion of the validity of the approach and the various assumptions in Section 5. The performance is compared to various other typical analysis methods in Section 6, while the computational burden is discussed in Section 7. Section 8 concludes with a brief summary of the article. Finally, in the appendices we provide a brief introduction to the NMO analysis itself (Appendix A) as well as details about the VLVνT toy model that we use to benchmark the performance of all considered analysis approaches (Appendix B).

2. Computational Challenge

The statistical comparison between experimental data and parametric or MC-based predictions allows inference of the values of physics parameters under study. As mentioned in the previous section, the comparison typically proceeds via a likelihood analysis. We first discuss its most general concepts and variations, then detail the arising computational requirements on MC generation, and finally present two standard methods of mitigating these computational burdens.

2.1. Likelihood Analysis

Different types of likelihood analyses in particle physics share common features[1]. An experiment records data which are used to reconstruct any observables expected to carry the imprint of the physical phenomenon under study. A selection (triggering, filtering, etc.) is applied in order to enhance the sought signal. Before performing statistical inference we need a theoretical model of the observable distributions to compare to the data. Often this includes complicated processes like particle interactions and detector response that require the use of MC methods. Hence, not only the data, but also the model is subject to statistical fluctuations. However, once an appropriate amount of MC events is available, the data x_i can be compared to templates—theoretical distributions—for different physics parameter values $\boldsymbol{\theta}$ via a likelihood function, $L(x_1, x_2, ..., x_n|\boldsymbol{\theta}) = \Pi_i P(x_i|\boldsymbol{\theta})$, where $P(x_i|\boldsymbol{\theta})$ is the probability to observe the data x_i assuming that $\boldsymbol{\theta}$ corresponds to given values of the physics parameters.

[1] See, for example, [4] for a more complete overview.

The goal is (in the frequentist picture) to find the maximum likelihood estimators (MLEs) $\hat{\boldsymbol{\theta}}$, i.e., the parameter values which maximize L.

A likelihood function can be applied to binned or to unbinned data. In the case of binned data one usually employs a Poisson likelihood or a χ^2 approximation thereof, the scenario on which the methods in this paper are based. Binning the data hides physics signatures smaller than the bin size and thus introduces a loss in sensitivity, which can be brought down to a negligible level by reducing bin sizes[2].

Apart from the physics parameter(s) of interest, a model often comes with nuisance parameters that are also included in the likelihood function. This further increases the dimensionality of the MLE search, which relies on numerical routines for multidimensional optimization problems. For the NMO studies, we use the L-BFGS-B algorithm [5] in a $D = 8$ dimensional parameter space (see Table A.4). This optimization process typically requires the generation of $\sim 10^3$ templates for a successful convergence.

2.2. Template and MC Generation Requirements

The problems associated with generating such a large number of templates are enhanced when estimating the median sensitivity of an experiment. The above process needs to be applied to an ensemble of pseudo-experiments[3] of size N_p. The comparison of the distributions of test statistics $\mathcal{T}$ (see Appendix A) can be used to estimate a significance value n_σ at which one hypothesis is preferred over the alternative. If $\mathcal{T}$ is Gaussian distributed[4], the uncertainty Δn_σ to which n_σ can be determined depends upon the number of pseudo-experiments N_p as $\Delta n_\sigma = \frac{1}{\sqrt{N_\mathrm{p}}}\sqrt{1+\frac{1}{n_\sigma^2}}$. With an absolute uncertainty Δn_σ at the 1% level, determining the sensitivity of an experiment at a confidence level of 99.7% (corresponding to $n_\sigma = 3$) requires on the order of 10^4 pseudo-experiments.

Taking a closer look at the sensitivity estimation in the two-hypotheses case of the NMO example reveals a further complication. In principle, for each pseudo-experiment generated under the true hypothesis one has to produce a complete distribution of $\mathcal{T}$ under the best fit $\hat{\boldsymbol{\theta}}$ in the opposite hypothesis in order to obtain one corresponding significance, and thereby build up the expected distribution of significance values. This makes the number of templates that need to be generated scale with N_p^2. The pragmatic solution to reduce this scaling behavior to N_p entails the assumption that the dependence of the test statistics distribution $\mathcal{T}$ on the parameters $\boldsymbol{\theta}$ is weak, so that $\mathcal{T}$ only needs to be calculated once for each ordering.

Finally, the event count expectation μ per bin in the templates must be determined at the same level of the physics effects being investigated, which requires at least $\frac{1}{(1\%)^2} = 10^4$ MC events per bin to study sub-percent variations arising in a comparison of the two NMO realizations. At the same time, the number of bins used in any histograms must

[2]While retaining sufficient MC statistics per bin, see discussion in Section 2.2.

[3]For certain problems, the generation of pseudo-experiments can be skipped by applying the *Asimov* approximation [6, 2].

[4]While not a prediction from the model, a near-Gaussian distribution of the test statistic is observed in most NMO studies [2, 3, 7].

be commensurate with the experimental resolution and the feature size of the effect under study. In the example case, at least $\sim \mathcal{O}(10^3)$ bins are required to resolve the distinct features of the NMO signature, otherwise the analysis cannot exploit the full potential of the experiment.

Therefore, in a brute-force approach, a very large number of neutrino events—in the example case $\sim \mathcal{O}(10^7)$—would need to be simulated for each of about 10^3 values of $\boldsymbol{\theta}$ probed during a single optimization process, for about 10^4 pseudo-experiments. Even if the time to simulate and reconstruct a single event is 1 s, full fits to all pseudo-experiments under the two ordering hypotheses would keep 10^5 CPU cores busy for 3 years to perform a single analysis[5]—a restriction clearly prohibitive to performing any study. Various methods can be employed to mitigate the high computational costs. These are very briefly discussed in the remainder of this section.

2.3. Weighting

The MC weighting technique avoids repeated simulation and reconstruction of events every time a value of a nuisance parameter is changed.

This is possible because the physics processes of initial neutrino production in the atmosphere, their propagation involving flavor oscillation, and their detection and reconstruction are independent. Only the event detection (plus subsequent reconstruction) exhibits a non-parameterizable dependence on its associated parameters, $\boldsymbol{\theta}_{\text{det}} \subseteq \boldsymbol{\theta}$, and therefore requires MC simulation. Each resulting MC neutrino of a given flavor β—generated for one unique realization of $\boldsymbol{\theta}_{\text{det}}$—is then assigned an individual weight w_β. This weight corresponds to the sum over the atmospheric fluxes Φ_α of all initial flavors α, including the probabilities $P^{\text{osc}}_{\alpha\to\beta}$ to oscillate into a neutrino of the flavor β:

$$w_\beta \propto \sum_\alpha \Phi_\alpha(\boldsymbol{\theta}_{\text{flux}}) \times P^{\text{osc}}_{\alpha\to\beta}(\boldsymbol{\theta}_{\text{osc}}).$$

Since the oscillation calculation is now decoupled from the detector simulation, only a single MC set is required to generate the templates of the different hypotheses under test (e.g., the two mass orderings). This eliminates statistical fluctuations between the otherwise disjoint MC samples.

However, even with a single MC set, an undersampling of the phase space of the model can result in a bias. In Section 6, for the example of the NMO analysis, the magnitude of this effect will be quantified together with a demonstration of the benefits of using our proposed method (*staged approach*).

2.4. Smoothing

Smoothing of the event distributions is a common practice when dealing with low statistics MC samples. For the various smoothing techniques—one of which is kernel density estimation (KDE) [8]—care has to be taken to not introduce artificial features which may

[5]Here we make the assumption that the algorithm can be parallelized perfectly.

incorrectly reduce or enhance the signal. To illustrate the performance of such a smoothing method, and to compare our method against, we use an adaptive bandwidth KDE directly on weighted MC. Here, a Gaussian kernel with a width calculated as described in [9] is centered at each MC event's reconstruction information. A weighted sum over the kernels of all events then delivers the smoothed distribution.

In the example NMO analysis presented in Section 6, we find that such a smoothing of the templates via KDE helps reduce biases, but not enough to make it viable with VLVνTs. Moreover, this method is impractically slow for our kind of analysis.

Shortcomings of the direct application of the two techniques discussed above—the first is the weighting method alone (labeled *direct histogramming*), while the second applies additional smoothing using adaptive kernel density estimates (labeled *direct KDE*)—can be overcome using the *staged approach*. Since it implements a combination of Monte Carlo and numerical integration methods[6], we first contrast the principles behind these, before we introduce the approach itself in the next section.

2.5. Integration Methods

Binning MC events in some observable dimension(s) according to their associated weights at the parameter values $\boldsymbol{\theta}$ corresponds to performing MC integration of the event distribution, whether it is smooth or has discontinuities. The result of this integration is an estimate of the expected event rate/count for any bin, $\mu(\boldsymbol{\theta})$. Errors of these estimates scale as $1/\sqrt{N}$, where N is the number of MC events that fall into the bin. Furthermore, the convergence of MC integration is independent of the dimension of the integral. Despite the presence of various techniques for improving convergence—most of which correspond to a clever sampling of the integration points (MC events)—huge computational efforts are often involved in obtaining a satisfying accuracy of the MC integration process.

The fact that physics processes that constitute a neutrino oscillation experiment can be grouped into independent effects and that some of these can be quantified without resorting to MC (see Section 2.3) makes numerical integration a promising alternative for the template generation process. For fixed computational cost and low dimensionality d, numerical integration performs favorably compared to MC integration: the uncertainty from simple trapezoidal integration, for example, is found to scale as $1/N^{-2/d}$ for large N [10], demonstrating that its converge rate exceeds that of MC integration in less than four dimensions.

3. Overview of the Staged Approach

The method to obtain templates we describe in this paper is divided into four independent parts, referred to as *stages*. The four stages (flux, oscillation, detection, and reconstruction) and how they are used to obtain event templates are summarized in this section, while more technical descriptions of each stage follow in Section 4.

[6] A review of these can be found in [10].

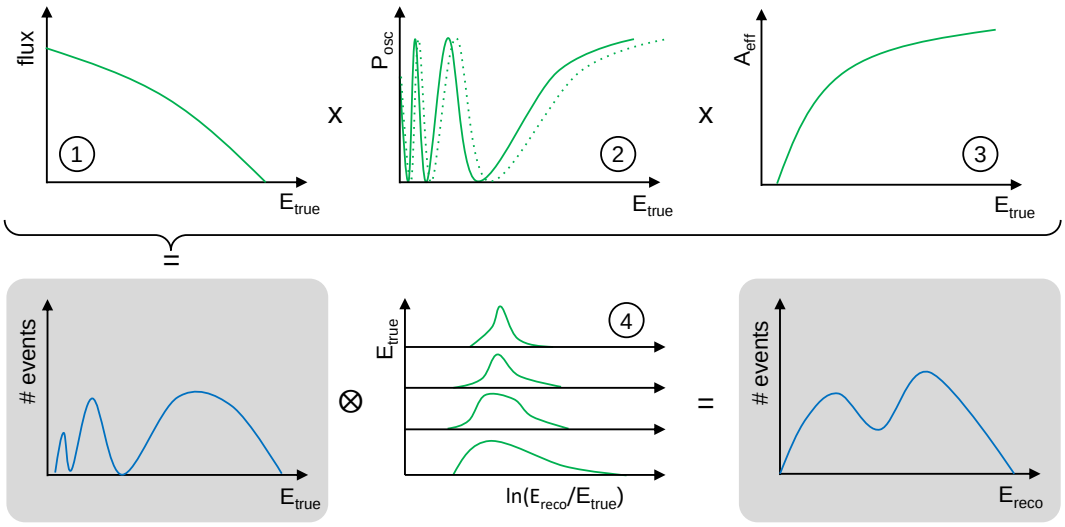


Figure 1: Illustration of the staged approach for obtaining event templates, here for simplicity using a characterization in one dimension (energy) only. Steps 1, 2, and 3 are in true energy (E_{true}); the product of these yields the expected event distribution (lower left). Smearing this spectrum with energy-dependent energy resolution functions (step 4) gives the reconstructed event rate spectrum (lower right). Note that the dotted green line in step 2 shows a hypothetical change of oscillation parameters, affecting only stage 2.

1. **Flux**
   The expected unoscillated atmospheric neutrino fluxes are taken from an external model [11]. Flux values from this model are provided in the form of tables with discrete steps in both neutrino energy, E_{true}[7], and direction, here the cosine of the zenith angle, $\cos\vartheta_{\text{true}}$. Therefore, an interpolation must be performed for values between those tabulated. Crucially, these tables give the integrated flux across the bins, which does not necessarily coincide with the flux value at the bin center. Accordingly, we use an integral-preserving (IP) interpolation. In general, atmospheric neutrino models require external inputs including primary cosmic ray measurements, atmospheric density models, and hadronic interaction measurements. Many associated uncertainties are known [12, 13] and need to be included as systematic parameters in an analysis.
2. **Oscillation**
   Flavor oscillations of neutrinos traversing the Earth modify the flavor content of the original flux in a manner that depends on the energies and path lengths (represented by the direction) of the neutrinos. Additional intrinsic neutrino properties describe the standard flavor oscillation probabilities: three mixing angles and two independent mass-squared splittings, as well as a possible non-zero CP-violating phase. In addition, matter effects can lead to an enhancement or deficit of oscillations compared to vacuum [14, 15, 16], which makes up the basis of the NMO measurement capability

[7]The use of the subscript "true" is used to specify the true variables of the neutrinos and to distinguish these from the reconstructed variables, denoted with a subscript "reco", which will be introduced later in this section.

of VLVνTs. In [16], the authors present an analytical expression for the neutrino flavor transition amplitude in a layer of uniform-density matter, which in turn was later implemented in the `Prob3++` software [17]. Here, the Earth density profile [18] is approximated by a finite number of homogeneous layers and the total transition amplitude is represented by a matrix product of the amplitudes in the individual layers. The main challenge for this stage is to keep the burden of these computationally expensive calculations to a minimum, while retaining sufficient accuracy in the modeling of the neutrinos' propagation.

3. **Detection**
   The number of actual events is the combination of the flux as well as a quantity known as the *effective area* (alternatively, the effective mass). This incorporates the probability that a given neutrino interacts within the detector, is detected, and passes the given data selection criteria. We obtain the eight effective areas ($\nu_{e,\mu,\tau}$ & $\bar{\nu}_{e,\mu,\tau}$ charged current (CC) and ν & $\bar{\nu}$ neutral current (NC) interactions) from simulated MC events that are run through the same selection criteria as the actual data. In general, each of these effective areas will depend on the energy and arrival direction of the neutrinos. Depending on the detector geometry, certain symmetries can be exploited to reduce the number of parameters which the effective areas are functions of. Here we assume azimuthal symmetry and therefore only extract effective areas as a function of E_{true} and $\cos\vartheta_{\text{true}}$.
4. **Reconstruction**
   The process referred to as *reconstruction* translates the raw signals recorded by a detector into physical properties of events, albeit imperfectly. How well these properties are estimated for the various neutrino flavors and interaction types can be seen as statistical distributions described by probability density functions, which we refer to as *resolution functions*. We estimate the resolution functions from MC events for which we know the true energy, zenith angle, and interaction type. The reconstruction stage uses these estimated resolution functions to build smearing kernels (ensembles of resolution functions) that map the event rates from the space of true variables into the space of reconstructed observables. Additionally—since most VLVνTs cannot exactly distinguish the different neutrino flavors and interaction types—the events are classified by their signature in the detector. Here, event classes are *tracks* and *cascades*, based on whether the event seems to contain the expected signature of a starting muon track. This process will separate ν_μ CC and $\bar{\nu}_\mu$ CC interactions from all others, albeit with limited efficiency and purity. For the example NMO analysis, three observables are needed: the primary neutrino's reconstructed energy (E_{reco}), zenith angle (ϑ_{reco}), and event classification.

In order to produce the final-level event templates that are ultimately compared to the data, the four stages are combined as depicted in Figure 1: integration of the product of the first three stages (flux, oscillation probability, and effective area) over E_{true} and $\cos\vartheta_{\text{true}}$ yields the event rate in the space of true variables. The event rate in the space of reconstructed observables is then obtained by a convolution of the true event rate with the

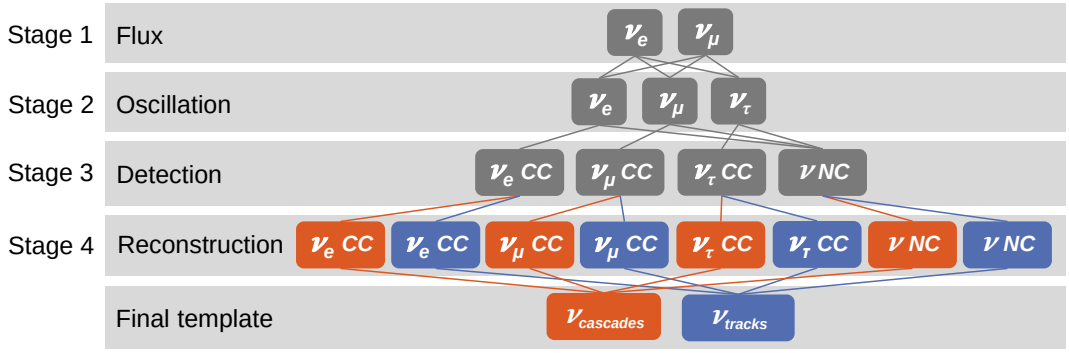


Figure 2: Flow of neutrino flavors and interaction types through the stages, here shown for neutrinos only (with an analogue counterpart for anti-neutrinos). Neutral current events of all flavors are indistinguishable and can therefore be conveniently added together. The reconstruction stage not only maps from ($E_{\text{true}} \times \cos\vartheta_{\text{true}}$)-space to ($E_{\text{reco}} \times \cos\vartheta_{\text{reco}}$)-space, but also classifies the events into the cascade and track categories, indicated by the orange and blue color, respectively.

reconstruction resolution kernels. Finally, multiplication by detector exposure time results in an event *count*, which can be compared directly to observed data or different templates[8].

Throughout the stages, different combinations of neutrino flavor and interaction type (channels) need to be treated separately, as depicted in Figure 2. Starting with the atmospheric flux, the neutrinos can undergo flavor change via oscillation. Since ν_τ production in the atmosphere is negligible at the energies relevant here, this flavor only appears through oscillation. The detection rate varies between between CC and NC interactions. Finally, after applying the reconstruction resolutions and event classification, event counts are summed to get the final-level templates for events classified as tracks and cascades separately. Where not mentioned explicitly, the same treatment is also applied to anti-neutrinos. The final templates are the sum over both, neutrinos and anti-neutrinos.

4. Technical Implementation of Stages

The stages within our approach, as summarized in Section 3 and illustrated in Figure 1, share some common properties but are also subject to different technical and computational challenges due to the physics effects each one captures. In this section we examine both generic and some specific implementation details which highlight how each stage balances performance and precision requirements—even in the presence of low MC statistics.

4.1. Common Features

Each stage represents an independent part of the experiment (i.e., a collection of related physical effects). With the exception of the (initial) flux stage, each subsequent stage applies a transformation to the output of the previous stage. In general, external information is

[8]While not shown here, it is possible to extend the model with more parameters or stages to describe additional effects, such as the modeling of systematic uncertainties.

required by each stage in order to do so. This can consist of a set of parameter values that are used to evaluate functional transformations, of external data, or of dedicated MC. Since the individual stages are independent of one another, a parameter change affecting one stage does not affect the transformation(s) used by the other three stages. Therefore, we include caching functionality that reduces the overall computational expense when a number of subsequent templates are retrieved while changing one parameter at a time.

The more basic mode of operation in which a stage's transformation is given by the actual parametric functions of the toy model itself (defined in Appendix B) is what we refer to as "truth". When relying on MC, events can simply be histogrammed with weights and in the dimensions relevant to the stage (MC integration within the stage). In Section 5.1 we employ this technique together with a high-statistics MC event sample to demonstrate the general validity of splitting up the template generation process into stages. Instead of simple histogramming, we can also apply smoothing methods to the transformations based on MC events in order to alleviate imprecisions due to statistical fluctuations. We validate the performance of said smoothing methods by sampling fixed numbers of toy MC events from the parametric functions, and passing these to the detection and reconstruction stages (Section 5.2).

To obtain the final outputs, first we define a binning in the space of E_{true} and $\cos\vartheta_{\text{true}}$, then calculate the flux at the bin centers, multiply with oscillation probability and effective area, and multiply by bin areas to obtain the event rates in the true variables. A convolution with the reconstruction resolution kernel is carried out as a discrete transformation between the same binning of true variables and the desired analysis binning in E_{reco} and $\cos\vartheta_{\text{reco}}$, separated into the event classifications. The choice of bins in each stage (for input, transformation, and output) is adjusted to reduce numerical integration errors and to avoid smearing out the physical effects under study. At the same time, the number of bins should be kept as small as possible to reduce the computational load. The same binning in true variables is used for the flux, oscillation, and detection stage, while reconstruction uses a coarser binning due to limited resolutions (see Section 5 for more details).

4.2. Flux

In order to preserve the integral of a tabulated set of data, a spline is fit to the *integral* of the data rather than to the values themselves. Interpolated values in the initial space are then found by evaluating the derivative of these splines. We call this method integral-preserving (IP) interpolation.

For the NMO example analysis, the tabulated data of interest are the atmospheric neutrino flux predictions from [11] provided as a function of both E_{true} and $\cos\vartheta_{\text{true}}$. To simplify the problem, the integration[9] is performed along one dimension at a time.

Consider the case with fluxes tabulated at $M \times N$ points in $(E_{\text{true}}, \cos\vartheta_{\text{true}})$. To retrieve the flux at an arbitrary $(E_{\text{true}}, \cos\vartheta_{\text{true}})$ point, say (x, y), first one spline of the conditional, integrated flux as a function of $\cos\vartheta_{\text{true}}$ is created for each of the M E_{true} locations. The derivative of each of these splines is evaluated at y, yielding M flux values. The integral

[9]Here, a cumulative sum of the bin values multiplied by the respective bin width.

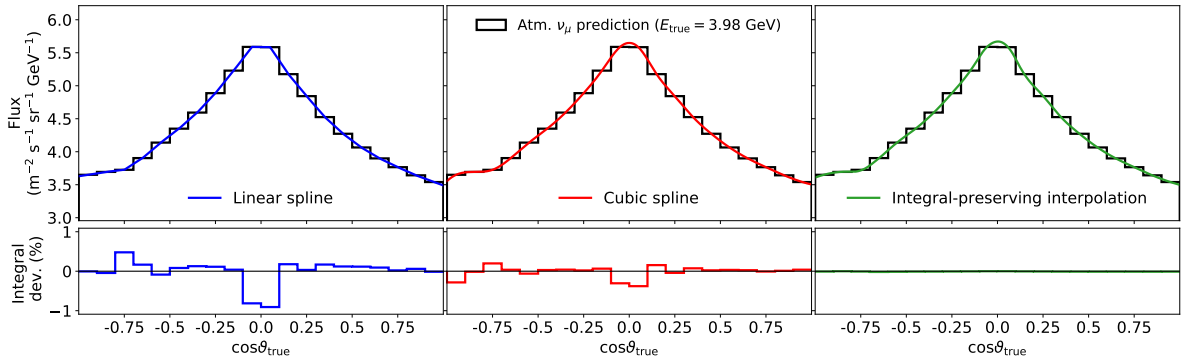


Figure 3: The top part of the figure shows three different interpolation methods applied to the same set of data points from [11] while the bottom portion shows the fractional deviation of the integral (= area under the curve) from these three methods. The deviations from the integral-preserving method presented in this paper have a maximum ∼0.02%.

of these values is then interpolated with a spline, and finally this spline's derivative is evaluated at x. This concept generalizes to higher dimensions, but can quickly become computationally intensive as the number of splines grows. For the example analysis of this paper, IP interpolation is approximately an order of magnitude slower than 2D cubic spline interpolation.

The IP method improves upon standard interpolation techniques in that it correctly models the turnover of the flux at the horizon ($\cos\vartheta_{\text{true}} = 0$) and the behavior in the most upgoing and downgoing regions ($\cos\vartheta_{\text{true}} \sim \pm 1$). This can be seen in Figure 3, which compares the results of IP to linear and cubic spline interpolation.

For the tables used in this paper's example analysis, IP interpolation preserves the integral to better than 0.5% over the complete (E_{true}, $\cos\vartheta_{\text{true}}$)-space.

4.3. Oscillation

The oscillation library that we employ is an extension of the code described in [19], where the authors ported some of the core functions of `Prob3++` to a graphics processing unit (GPU) via the CUDA C API. We then added back in the ability to handle an arbitrary number of constant density layers of matter, allowing for highly parallel calculations of three-flavor oscillation probabilities of neutrinos that encounter a realistic radial Earth density profile, with fine-grained control over its characteristics. We implemented the oscillation calculations with floating point precision selectable to either single (32 bits, or FP32) or double (64 bits, or FP64) precision. With our code run in double precision with `Prob3++`, evaluated on a 100×100 grid of neutrino energies E_{true} ranging from 1 GeV to 80 GeV and $\cos\vartheta_{\text{true}}$ values spanning the upgoing region, our GPU and CPU implementations of the `Prob3++` code produce consistent results to the level of 10^{-14} or less. These differences are due to differing hardware implementations of the same mathematical operations. Switching from double to single precision on the GPU, we find that the magnitudes of the differences stay below about 10^{-5} for all oscillation channels. Single precision is desirable from a performance point of view, since most GPUs comprise a larger number of single precision than double precision

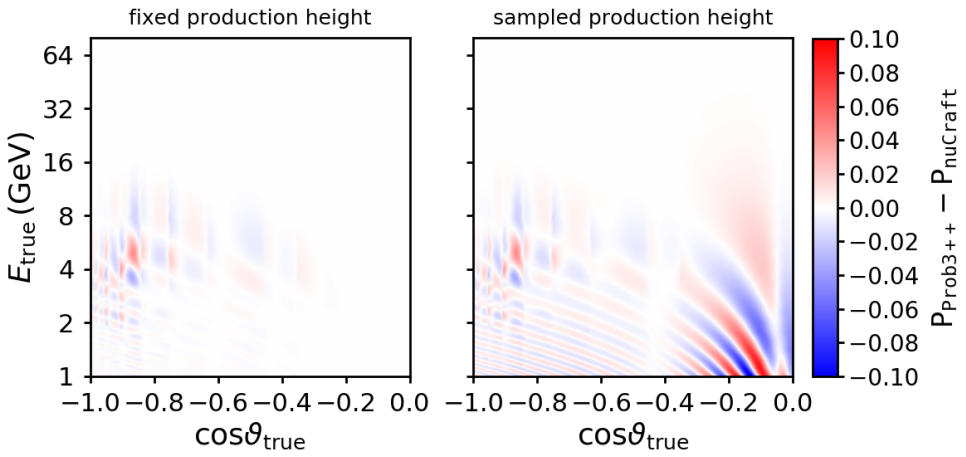


Figure 4: Deviation of ν_μ survival probabilities computed with `Prob3++` compared to `nuCraft`. The left panel uses a fixed production height of 20 km for both codes and twelve constant-density layers for `Prob3++`. In the right panel the values from `nuCraft` are the average probabilities for a range of neutrino production heights across the atmosphere.

arithmetic units, and these extra units can be exploited by the parallelism in our code.

To evaluate the effects of an approximated Earth density profile using a limited number of constant density layers and a constant atmospheric production height—both approximations that our code makes—we compare the oscillation probabilities from our implementation of `Prob3++` against a reference model. The latter is calculated by `nuCraft` [20], which is written in Python and solves the Schrödinger equation numerically. The `nuCraft` library also supports a realistic variation of the oscillation baselines according to the distribution of atmospheric neutrino production heights described in [21] and uses an interpolated radial density profile of the Earth.

To this effect, we first fix the atmospheric neutrino production height to $h_0 = 20\,\text{km}$ for both codes, and quantify the deviations arising from the coarser Earth model by calculating the ν_μ survival probability residuals on a fine grid in cosine zenith and energy. When approximating the Earth's density profile with only four layers (one for each of the upper and lower mantle, and the outer and inner core), differences of up to 15 % to the output of `nuCraft` are seen. These differences decrease to below 5 % when using 12 density layers (see left panel of Figure 4). Using an even more detailed model with 59 layers results in differences smaller than 0.3 % across the whole two-dimensional spectrum.

Comparing the 12-layer `Prob3++` probabilities to those obtained under the assumption of a more realistic distributed atmospheric production height in `nuCraft` highlights further discrepancies between the outputs of the two codes (see right panel of Figure 4). However, the largest differences ($\sim \pm 10\,\%$) appear for near horizontal trajectories, while the residuals for $\cos\vartheta \lesssim -0.4$ remain roughly unchanged.

Since precise modeling of both the Earth's density profile and the atmospheric neutrino production heights come at a significant additional computational cost, depending on

the analysis in question it might be desirable (and justifiable) to neglect one or both of these effects. In our example NMO analysis we find that it is sufficient to use the 12-layer model and a fixed production height. Both approximations have very little impact on the final spectra—mainly due to detector resolution effects—and since they systematically affect both NMO realizations in an almost identical manner, their effects largely cancel out in a comparison of the two. Moreover, while the atmospheric flux peaks in horizontal direction (seen, for example, in Figure 5), negligible matter effects for the corresponding trajectories result in very little intrinsic sensitivity of this part of the spectrum to the NMO.

4.4. Detection

As a reminder, the effective areas are quantities used to translate an incoming flux to the event rates in the detector. These effective areas are calculated from a limited number of MC events, hence they can suffer from statistical fluctuations which can be a non-negligible contribution to the total uncertainty of the final physics result. At the same time, effective areas are typically well-behaved quantities in energy and zenith angle (under some realistic assumptions, e.g., that no discontinuous selection cuts are applied and no gaps exist in the detector acceptance). Therefore, smoothing techniques can be applied to alleviate the unwanted uncertainty contributions from statistical fluctuations.

For charged current interactions, we compute the effective area separately for each neutrino flavor. In contrast, we do not distinguish between flavors for neutral current (NC) interactions, since their cross sections are identical. Neutrinos and anti-neutrinos are handled independently, accounting for a total of eight independent effective area functions. For convenience we include the multiplication by detector exposure time (t_{exp}) in the same step, which means that this stage outputs event counts (N_{events}) instead of rates

$$N_{\text{events}} = \Phi[\text{m}^{-2}\text{s}^{-1}] \cdot A_{\text{eff}}[\text{m}^2] \cdot t_{\text{exp}}[\text{s}]\,, \tag{1}$$

for some input flux (Φ).

In our staged approach we first evaluate the effective areas on a fine grid in (E_{true}, $\cos\vartheta_{\text{true}}$) using the MC events via MC integration. For small sample sizes, some grid points may have no associated events, leading to gaps in the distribution. We remove these by applying a simple Gaussian smearing along the two-dimensional grid. In a second step, cubic splines are employed to perform smoothing regression in the E_{true} and $\cos\vartheta_{\text{true}}$ dimensions.

Figure 5 shows the true template of ν_μ CC events (obtained by operating the detection stage in parametric mode) together with the deviations that arise when the same template is obtained from MC samples[10] of different sizes using direct histogramming and our proposed smoothing method. We use ν_μ CC events as an example here and below. Table 1 gives the average (binwise) and maximal χ^2 values by which the templates from the proposed method and from direct histogramming deviate from the true templates. Applying our method we find deviations that are lower by a factor of about 40 for the smallest MC set, and by a factor of about 13 for the largest. It is noteworthy that the maximum deviation (χ^2_{max}) across all bins decreases monotonically with MC sample size, confirming that the proposed method does not introduce any observable bias.

[10]Generated from the toy model in Appendix B.

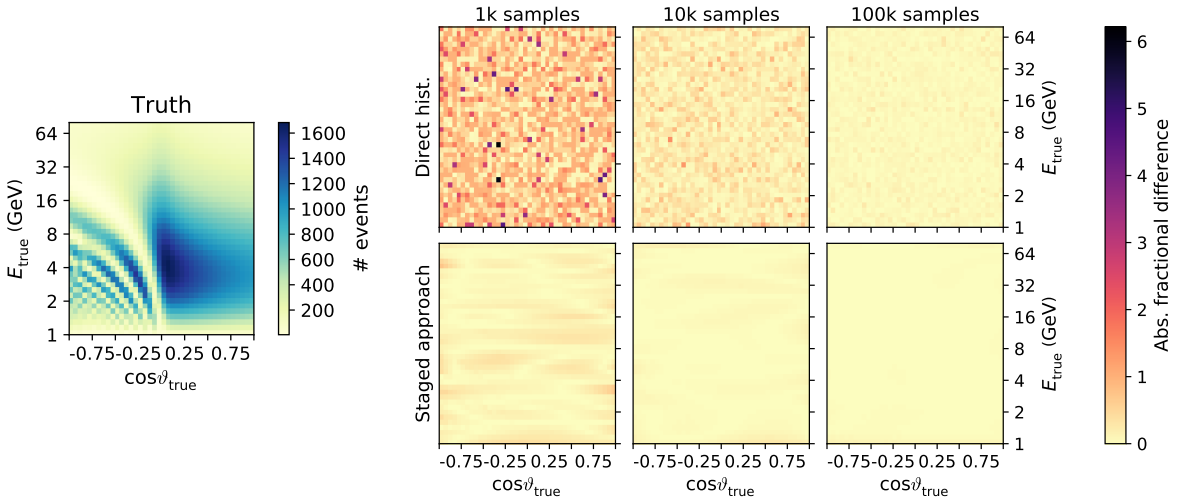


Figure 5: Parametric reference distribution after the first three stages (flux, oscillation, and detection) for the ν_μ CC channel in ($\cos\vartheta_{\text{true}}$, E_{true}) (left panel) and relative residuals for the direct histogramming (right panel, top row) and our proposed method (right panel, bottom row) using different amounts of simulated events. The three columns in the right panel represent different MC sample sizes of 10^3, 10^4, and 10^5 events, respectively. The samples are drawn from the unbinned toy model distributions of Appendix B.

4.5. Reconstruction

The usual way to obtain templates in the space of reconstructed variables is to place each individual MC event according to the reconstruction information that the event carries. This is the case for both methods that are used for comparison: direct histogramming and direct KDE. In contrast, the staged approach follows a different prescription that is not based on single event information. Instead, the available MC simulation is used to construct detector resolution functions to map a template in true variables (such as in Figure 5) into its counterpart in the space of reconstructed variables.

In the case study of the NMO analysis, the mapping of true variables (E_{true} and $\cos\vartheta_{\text{true}}$) to reconstructed variables (E_{reco}, $\cos\vartheta_{\text{reco}}$, and event classification) can be extracted from

Sample size		10^3	10^4	10^5	10^6
Direct hist.	$\langle\chi^2\rangle$	215	22.5	2.07	0.201
	χ^2_{max}	21600	1810	79.4	11.2
Staged approach	$\langle\chi^2\rangle$	5.14	0.526	0.0615	0.0156
	χ^2_{max}	460	17.2	2.27	0.975

Table 1: Average χ^2 across all bins and the worst-case bin's χ^2 value comparing templates in (E_{true}, $\cos\vartheta_{\text{true}}$)-space (i.e., before applying reconstruction resolutions) generated by direct histogramming (top) and the smoothed-staged approach (bottom) with the toy model's reference template. Shown are values obtained for independent input MC samples of various sizes (from 10^3 up to 10^6 events per flavor/interaction type).

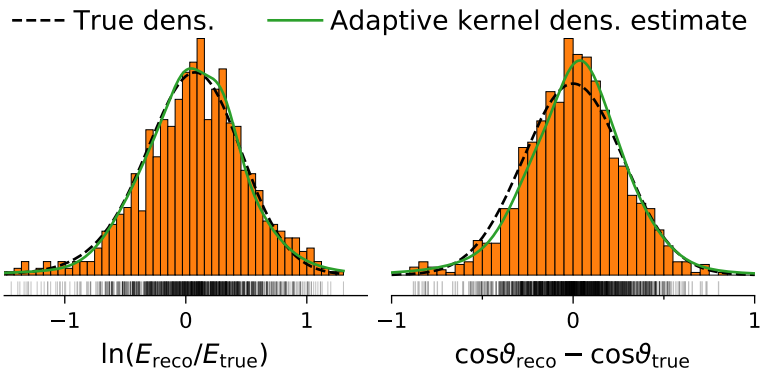


Figure 6: Example energy and cosine-zenith-angle resolution distributions for ν_μ CC events classified as cascades, estimated with histograms and adaptive KDE. Energy resolution is shown for 100 events with $E_{\text{true}} \in [26.7, 29.8]\,\text{GeV}$ and cosine-zenith resolution for 100 events with $E_{\text{true}} \in [1.0, 1.1]\,\text{GeV}$. The samples used to construct the histogram and KDE are shown by vertical lines beneath the histograms.

the MC as a five-dimensional linear transform. Due to the high dimensionality, the reconstruction stage is particularly sensitive to small MC sample sizes: for a required number of n events per bin, it would require $\mathcal{O}(n^5)$ MC events to generate separate resolution kernels for each E_{true} and $\cos\vartheta_{\text{true}}$ value[11].

If the functional form of the resolution functions is known, a parametric model fit to the MC yields the most accurate estimate. Utilizing such a parametric model produces the most robust templates. However, as we do not know the form of these functions, a non-parametric density estimation technique is used to approximate them. In particular, we chose to use adaptive KDE [22] since it is well suited for the non-trivial distributions encountered in real detector simulations. To compute a resolution function (e.g., $\ln(E_{\text{reco}}/E_{\text{true}})$) using a set of MC events, we can estimate the true distribution via KDE. This is done by placing a kernel function (here a simple Gaussian) centered at each event's value of the variable to be described and then summing over all kernels. In contrast to more commonly used fixed-bandwidth versions, an adaptive KDE modifies the width of each kernel. In our example these bandwidths are chosen via the improved Sheather Jones (ISJ) algorithm [23], which does not assume that data is normally distributed, in contrast to predecessor algorithms (see also [24, 9]). An example of two resolution functions (one for both energy and zenith angle, respectively) estimated using the adaptive KDE method is shown in Figure 6.

Furthermore, our resolution functions can be constructed in a way that they apply to a large fraction of events at the same time, which is useful to minimize statistical fluctuations. Characterizing the expression $\ln(E_{\text{reco}}/E_{\text{true}})$ instead of E_{reco} reduces dependency on E_{true} and makes this a relatively slowly changing quantity. This is in contrast to the previous stages which typically deal with faster changing features such as the rapidly varying oscillation weights or effective areas near the energy thresholds. Thus, more events can be

[11] This differs from the detection stage which is also MC-based but is a mapping that involves only two dimensions and hence requires a more manageable $\mathcal{O}(n^2)$ MC events to meet the requirement.

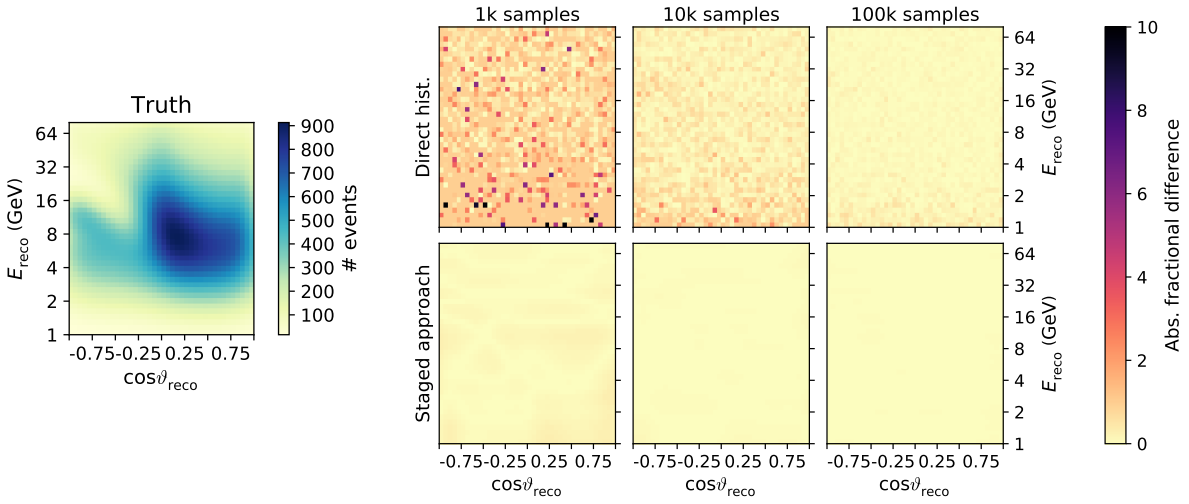


Figure 7: Same as Figure 5, but comparing final-level templates after all four stages are applied. Note that the residuals in the 1k-samples plot for direct histogramming go up to 31 but the scale is clipped at 10.

combined and treated together here.

For the resolution of the neutrino zenith angle we similarly characterize $\cos\vartheta_{\mathrm{reco}} - \cos\vartheta_{\mathrm{true}}$. A direct mapping is used for event classification. While the latter is mostly independent of the neutrino zenith angle $\cos\vartheta_{\mathrm{true}}$ in a real detector (and for our toy detector model, it is independent by construction) and thus many events can be grouped together, the angular resolutions have more complicated dependencies—on zenith angle, energy, and event signature—and are treated in a fully differential way. (In the case of our simple toy detector model, however, we assume them to be independent of zenith angle.) The resolution functions defined in this way change only slowly as a function of E_{true} and $\cos\vartheta_{\mathrm{true}}$.

Figure 7 again demonstrates that templates obtained from our KDE based reconstruction stage deviate much less from the parametric reference template after reconstruction than templates from direct histogramming of reconstructed MC events.

5. Validation and Comparison of Templates

This section more closely examines the templates generated with the staged approach and compares them—along with those generated by the other two methods (histograms and KDE)—to the parametric reference distributions of the toy detector model. This validation is split into two parts. The first examines the grid of points that are used to numerically approximate the integral over the first three stages, whereas the effect of smoothing is investigated in the second.

5.1. Sampling Grid

To keep the computational burden low, we evaluate all stages on a fixed grid of points distributed over E_{true} and $\cos\vartheta_{\mathrm{true}}$, while we output the final templates with a binning of $40\times$

Grid ($M \times N$)	40×40	80×80	160×160	320×320	640×640	1280×1280
$\langle \chi^2 \rangle$	0.01067	0.00253	0.00060	0.00014	0.00003	0.00001
χ^2_{max}	1.45906	0.46930	0.19718	0.04974	0.00634	0.00172

Table 2: Average and maximal χ^2 deviations between final templates of non-smoothed staged approach and direct histogramming, for different grid point densities in (E_{true}, $\cos\vartheta_{\text{true}}$) for the first three stages, using an MC sample of 10^6 events. The last (=reconstruction) stage uses a reduced binning, as described in the text.

40×2 in E_{reco}, $\cos\vartheta_{\text{reco}}$, and event type (determined by the scales of the physics signatures). By comparing the staged approach without smoothing to histograms, we demonstrate the validity of our technique, and show that it becomes equivalent to traditional MC weighting as grid point spacing in E_{true} and $\cos\vartheta_{\text{true}}$ is reduced.

For the staged approach we use a grid of N equally spaced $\cos\vartheta_{\text{true}}$ points between -1 and 1, and M logarithmically spaced points in E_{true} ranging from 1 GeV to 80 GeV. The $\cos\vartheta_{\text{true}}$ values of the simulated events are generated randomly following a uniform distribution, while the neutrino energies are drawn from a power law $\propto E_{\text{true}}^{-1}$.

Table 2 shows the χ^2 difference between the final templates obtained from direct histogramming and the staged approach for a variety of grid point densities in E_{true} and $\cos\vartheta_{\text{true}}$, using the same MC set of size 10^6 for both methods. The relative decrease in the average χ^2 value roughly scales with the inverse of the relative grid density increase, thus confirming that the two methods will agree to arbitrary precision in the asymptotic limit. In the following, for practical reasons we limit ourselves to a grid of 400×400 points in E_{true} and $\cos\vartheta_{\text{true}}$ (further reduced to 200×200 for the last (=reconstruction) stage).

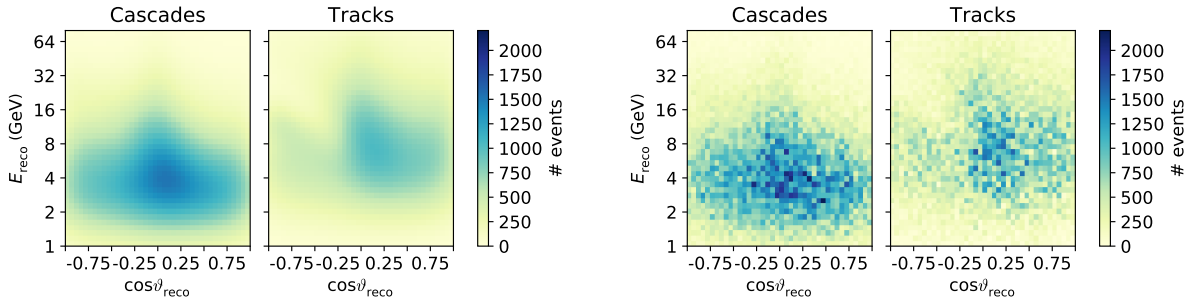
5.2. Smoothing

To validate the final templates with smoothing applied at each stage, we compare them directly to truth. For reference, we also show the agreement resulting from both the direct histogramming and the direct KDE methods.

While Table 3 quantifies deviations from the reference distributions again in terms of χ^2 and in dependence of MC sample size, Figure 8 displays the final-level templates for each of the aforementioned methods using a sample with 10^4 events.

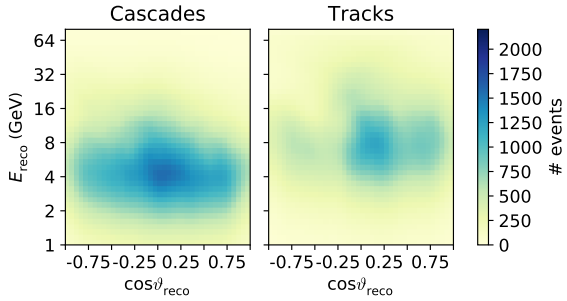
The staged approach outperforms the two alternatives in terms of χ^2 values by more than one order of magnitude for all those sample sizes studied here. Furthermore, inaccuracies of the templates from the staged approach scale with the inverse of sample size almost as fast as those of templates from direct histogramming. In addition, it is noteworthy that the KDE method shows comparably slow convergence, i.e., it performs worse than direct histogramming for the sample size of 10^6.

Considering a sample size of 10^4 and the staged approach, the average χ^2 is only about 30% of what is expected just from statistical fluctuations in data (where a χ^2 of 1.0 is expected per bin), while more than 10^6 events would be necessary to achieve the same average χ^2 using direct histogramming or KDE. Therefore, to reach an equal accuracy, one to two orders of magnitude larger samples are needed for histogramming or KDE compared

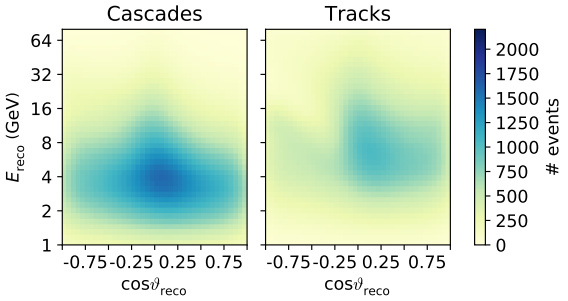


(a) Truth

(b) Direct Histogramming



(c) Direct KDE



(d) Staged Approach

Figure 8: Final-level templates used for the example data analysis. The reference distributions (truth) obtained directly for the toy detector model parameterizations are shown in panel (a). Given the same sample of 10^4 events the estimated distributions using histograms are shown in panel (b), using KDEs in panel (c), and using the staged approach in panel (d).

to the staged approach. The next section illustrates the implications for running a data analysis.

6. Example Analysis Results

To illustrate the impact of sample size, we show the resulting $\sqrt{\Delta\chi^2}$ as an estimate for the sensitivity to the NMO for our example analysis in Figure 9. For reference, the true result is derived directly from the exact templates based on the parametric toy detector model and lies at $\sqrt{\Delta\chi^2} = 5.75$. For the three methods discussed throughout this paper, the statistical uncertainty of the obtained sensitivity is indicated by error bars in the figure. This statistical uncertainty is computed from several statistically independent MC sets. These uncertainties reveal that the methods exhibit quite different intrinsic fluctuation of their respective sensitivity estimates, as well as different scaling behavior of the variance with sample size. As sample size decreases, direct histogramming without any smoothing applied results in an increasing overestimation of a VLVνT's ability to exclude the wrong neutrino mass ordering. In the most extreme case shown here (corresponding to the smallest sample size of 10^3), the sensitivity is estimated to be more than one order of magnitude greater than the actual capability of the experiment, mostly due to the bias from undersampling

Sample size		10^3	10^4	10^5	10^6
Direct Histogramming	$\langle\chi^2\rangle$	468	42.6	4.27	0.458
	$\chi^2_{\max}$	$3.4 \cdot 10^4$	906	138	10.5
Direct KDE	$\langle\chi^2\rangle$	32.2	11.4	3.67	1.25
	$\chi^2_{\max}$	245	90.2	50.3	25.3
Staged Approach	$\langle\chi^2\rangle$	3.01	0.303	0.111	0.0301
	$\chi^2_{\max}$	47.4	3.03	1.80	0.387

Table 3: Average and maximal χ^2 deviations between final templates of the three shown methods and truth, for independent input MC samples of various sizes. Note that the staged approach has smoothing applied (the default), in contrast to Table 2.

the reconstruction space.

Applying KDE smoothing to the weighted events instead of histogramming them (direct KDE) leads to a reduction of the overestimated sensitivity for sample sizes of up to at least $3 \cdot 10^5$ but is not able to eliminate the bias entirely for the tested sample sizes. For sample sizes larger than $\mathcal{O}(10^5)$, the direct KDE method is too computationally expensive to deliver results within a reasonable time (for more details on timing, see Section 7).

The estimated sensitivity using the staged approach is statistically compatible with the true sensitivity across the whole range of sample sizes considered. It shows no bias and lower variance for predicting sensitivity to physics compared to the other methods within the limits of our testing.

7. Benchmarks

Whether a given analysis method is useful in a realistic setting depends not only on its numerical reliability, but also on how long it takes to compute the quantity of interest (note that this duration is in addition to the initial time needed to generate the MC). For reference, we performed benchmarks of the template generation times in the course of a typical analysis process[12]. These are compiled in Figure 10.

Note that no initial start-up times—such as the construction of the smearing kernels used within the reconstruction stage—are included here. For all three methods separate timings based on our CPU-only and GPU-accelerated implementations are provided.

While for sample sizes below 10^4 to 10^5 events direct histogramming is the fastest method, it is unusable owing to the large fluctuations associated with the templates it produces, which in turn results in the grossly overestimated sensitivites shown in Figure 9. Direct KDE only proves competitive when used in conjunction with the smallest datasets. The faster-than-linear scaling of its computational needs with sample size then quickly renders it impractical to use. Our proposed method is independent of sample size by construction (excluding initial start-up costs), but will get more expensive if a finer grid point spacing is desired.

[12]Timings were obtained on a computer with an Intel Xeon E5-1660 v3 3.0 GHz CPU and an Nvidia GeForce GTX Titan X GPU.

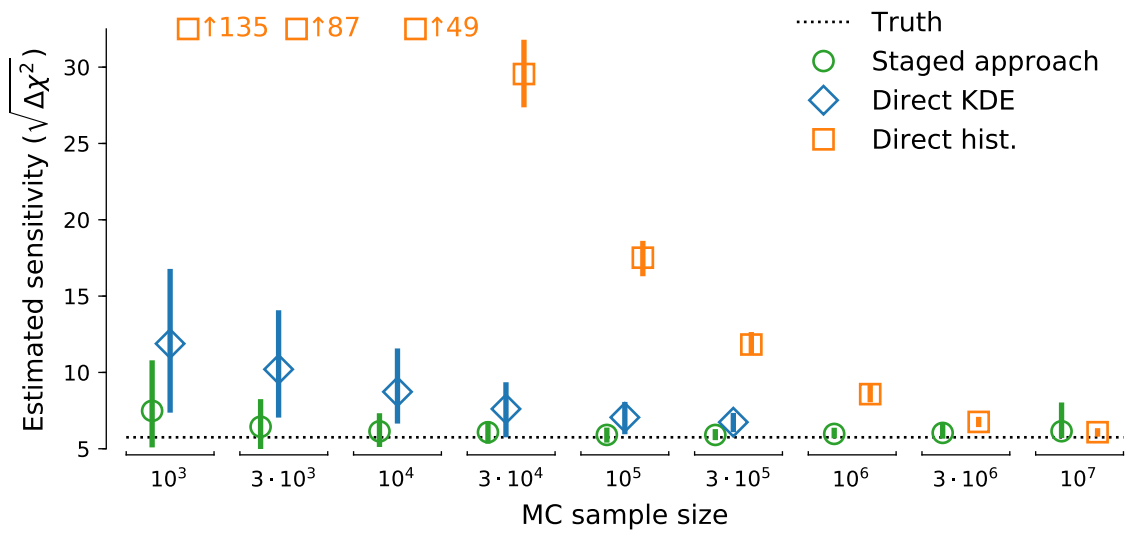


Figure 9: Estimated sensitivity ($\sqrt{\Delta\chi^2}$) to the NMO vs. sample size for direct histogramming, direct KDE, and the proposed staged smoothing methods applied to multiple (between 50 and 200) statistically independent toy MC sets. Vertical lines indicate central 68% quantiles. The dashed horizontal line shows the significance obtained from truth templates based on the parametric toy detector model. The staged approach outperforms the other methods—both in terms of bias and variance—for sample sizes through $3 \cdot 10^6$, with direct histogramming only matching the staged approach using 10^7 samples. Note that no data points exist for direct KDE and sample sizes above $3 \cdot 10^5$, as computational processing times become impractically large. Also note that direct histogramming is off-scale high for fewer than $3{\cdot}10^4$ events (mean values are indicated to the right of the corresponding markers).

The timing difference between the CPU and GPU implementation of the staged approach is not as large as for the other methods, since it is only using the GPU for parallelization of the neutrino oscillation weights calculation (whereas the other methods make use of the GPU more extensively).

8. Summary

The search for small physics effects in high statistics neutrino oscillation experiments requires careful treatment and use of simulated data. Statistical fluctuations within distributions obtained from Monte Carlo simulations are able to severely distort an analysis, rendering derived constraints or sensitivities essentially meaningless.

The staged approach we have presented serves two main purposes. Firstly, computational expense is reduced through sampling of physics and detector response distributions on a discrete grid instead of computing a weight for every individual Monte Carlo event. In this respect, we have demonstrated that our method of breaking down the template generation into independent stages converges to the MC weighting scheme when using a grid of a high enough, albeit feasible, density. For a fixed number of grid points, the template generation time has been shown to be independent of the input sample size. Secondly, the staged

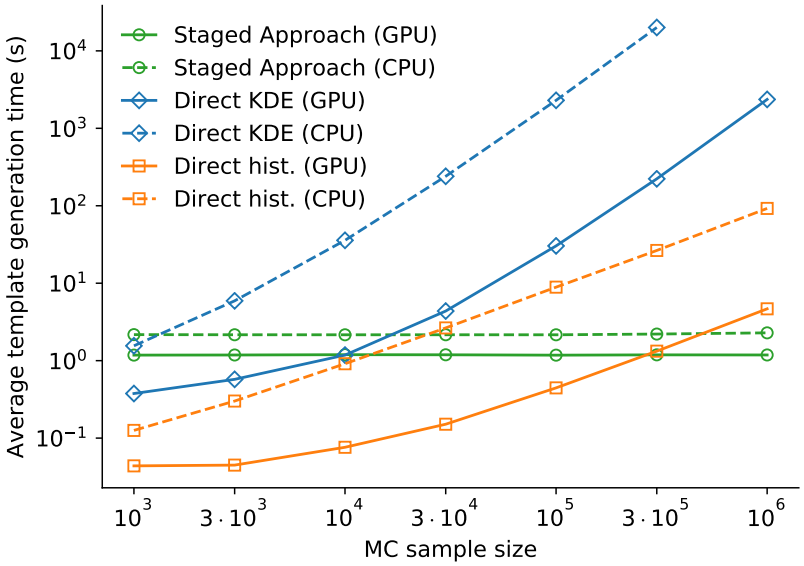


Figure 10: Average template generation time during a typical analysis for input datasets of varying size, shown for the direct histogramming, the direct KDE, and the staged approach. Solid lines represent timings based on (partial) GPU acceleration, whereas the dashed ones are for CPU-only calculations.

approach allows the application of smoothing techniques to a detector's response functions. This has proven superior to the smoothing of the final event distributions since it is faster and—even more importantly—yields more accurate and robust results.

In the example neutrino mass ordering analysis that we have conducted—to benchmark and compare the different approaches—we found that direct histogramming of events leads to a gross overestimation of sensitivities when used in conjunction with small ($\lesssim 10^6$ events for our toy model) numbers of samples. Conversely, the proposed staged approach leads to correct results that are largely unaffected by the sample size across the tested range and the variance of results is small compared to the result above about 10^4 neutrino events. This means that the necessary amount of simulated events is reduced significantly (by about two orders of magnitude in our example)—an important aspect especially since Monte Carlo event simulation and reconstruction times can represent major hurdles to progress in the field of neutrino oscillation experiments.

Acknowledgments

The authors gratefully acknowledge the support from the following agencies and institutions: USA – U.S. National Science Foundation-Office of Polar Programs, U.S. National Science Foundation-Physics Division, Wisconsin Alumni Research Foundation, Center for High Throughput Computing (CHTC) at the University of Wisconsin-Madison, Open Science Grid (OSG), Extreme Science and Engineering Discovery Environment (XSEDE), U.S. Department of Energy-National Energy Research Scientific Computing Center, Particle astrophysics research computing center at the University of Maryland, Institute for Cyber-

Enabled Research at Michigan State University, and Astroparticle physics computational facility at Marquette University; Belgium – Funds for Scientific Research (FRS-FNRS and FWO), FWO Odysseus and Big Science programmes, and Belgian Federal Science Policy Office (Belspo); Germany – Bundesministerium für Bildung und Forschung (BMBF), Deutsche Forschungsgemeinschaft (DFG), Helmholtz Alliance for Astroparticle Physics (HAP), Initiative and Networking Fund of the Helmholtz Association, Deutsches Elektronen Synchrotron (DESY), and High Performance Computing cluster of the RWTH Aachen; Sweden – Swedish Research Council, Swedish Polar Research Secretariat, Swedish National Infrastructure for Computing (SNIC), and Knut and Alice Wallenberg Foundation; Australia – Australian Research Council; Canada – Natural Sciences and Engineering Research Council of Canada, Calcul Québec, Compute Ontario, Canada Foundation for Innovation, WestGrid, and Compute Canada; Denmark – Villum Fonden, Danish National Research Foundation (DNRF); New Zealand – Marsden Fund; Japan – Japan Society for Promotion of Science (JSPS) and Institute for Global Prominent Research (IGPR) of Chiba University; Korea – National Research Foundation of Korea (NRF); Switzerland – Swiss National Science Foundation (SNSF); United Kingdom - Science and Technology Facilities Council (STFC).

Appendix A. NMO Analysis

The observation of neutrino oscillations and the demonstration of the neutrinos' non-zero masses [25, 26] represented a major step forward in the field of particle physics. While current experimental techniques have not yet allowed for a direct measurement of the tiny masses, the magnitudes of their relative differences (mass splittings) are well known.

The ordering of these neutrino mass states (neutrino mass ordering, NMO) presents a difficult challenge. A powerful method to determine this ordering is the observation of matter effects on neutrinos mentioned previously in Section 3. Owing to the high electron density of the Sun, observations of solar neutrinos have shown the second mass state to be heavier than the first [27]. It still remains an open question, however, whether the third state is the most or least massive. The former scenario is referred to as the normal ordering (NO), while the second is called inverted ordering (IO). There is currently no experimental evidence excluding either of the two scenarios.

The study of oscillations of atmospheric neutrinos provides a promising route toward a decisive measurement of the NMO [28, 1, 2, 3]. The path length (or *baseline*) varies between 20 km for vertically downward going and 12 700 km for straight upward going atmospheric neutrinos, with the latter crossing the full diameter of the Earth. With energies ranging from MeV up to the TeV scale, combinations of baselines and energies varying over several orders of magnitude are probed. For the longest baseline, the very pronounced first oscillation maximum of muon neutrinos occurs at a neutrino energy of around 25 GeV, followed by subsequent maxima at lower energies.

The electron neutrinos' coupling to electrons (coherent forward scattering) in the Earth creates an effective matter potential which leads to resonant behavior of the transition probabilities at energies around 5 GeV, known as matter resonances [15, 14, 29]. Furthermore, the Earth's specific density profile encountered by the neutrinos can also parametrically

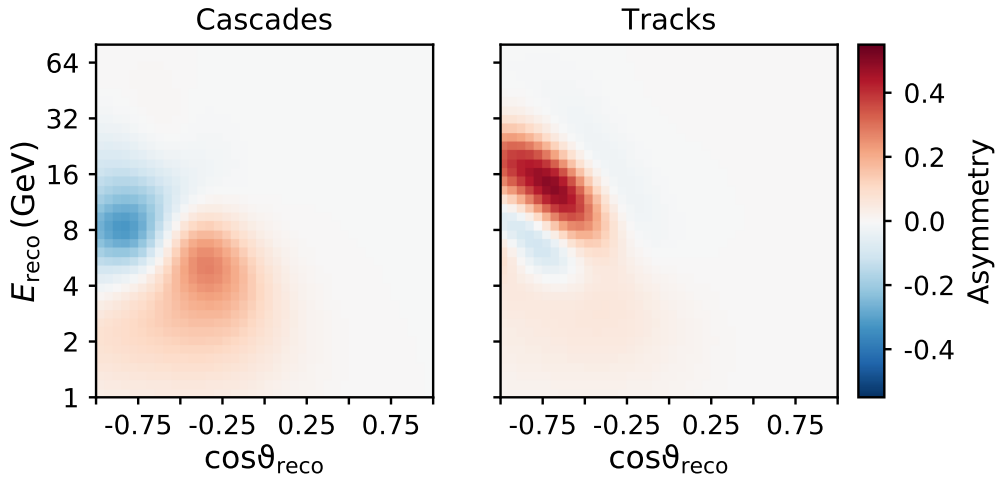


Figure A.11: Annual NMO asymmetry without systematics for the toy model, for cascades (left) and tracks (right). The asymmetry is defined as a bin-wise $\sqrt{\chi^2}$ between the IO and NO hypotheses (see Table A.4) and can thus be interpreted as an NMO sensitivity proxy in the absence of systematic error sources.

enhance their oscillations [30]. This enhancement with respect to oscillations proceeding in vacuum occurs for neutrinos if the NMO is normal, otherwise for anti-neutrinos.

The NMO measurement potential of VLVνTs is based on this asymmetry. Two major aspects are obstructive, however. The first is the inability of VLVνTs to differentiate between neutrinos and anti-neutrinos. This reduces the effect to the respective difference in atmospheric fluxes and interaction cross sections. Energy and directional resolutions of the experiment present the second hurdle. Both are typically prohibitive to resolving the fast variations of the oscillation pattern at the relevant energies. As a consequence, the observable effect is reduced to at most a few percent over the relevant energy and zenith range (see Figure A.11), requiring neutrino telescopes with effective masses on the order of megatons to achieve sufficient event statistics.

Proponents of various VLVνTs in ice and water have performed studies contributing to a solidification of this idea, finding that a $> 3\sigma$ (median) sensitivity to the NMO can be achieved within five years of exposure time even in less favorable regions of the neutrino oscillation parameter space [1, 3, 31].

As the oscillation probabilities directly depend on neutrino energy E_{true}, oscillation baseline ($\propto \cos\vartheta_{\mathrm{true}}$), and flavor, we split our data into bins of $\log_{10}E_{\mathrm{reco}}$, $\cos\vartheta_{\mathrm{reco}}$, and event class. It is important to choose a binning fine enough to resolve the NMO signature, while coarse enough to retain a sufficient amount of MC statistics per bin, as motivated in Section 2. We have found the division into $(40 \times 40 \times 2)$ bins to be suitable, covering a range of E_{reco} from 1 GeV to 80 GeV, the whole sky ($\cos\vartheta_{\mathrm{reco}}$ from -1 to 1), and the two event classes of *cascades* and *tracks*. Figure A.11 shows the bin-wise $\sqrt{\chi^2}$ difference between an inverted and a normal ordering spectrum using this binning, based on the two sets of nominal model parameter values given in Table A.4.

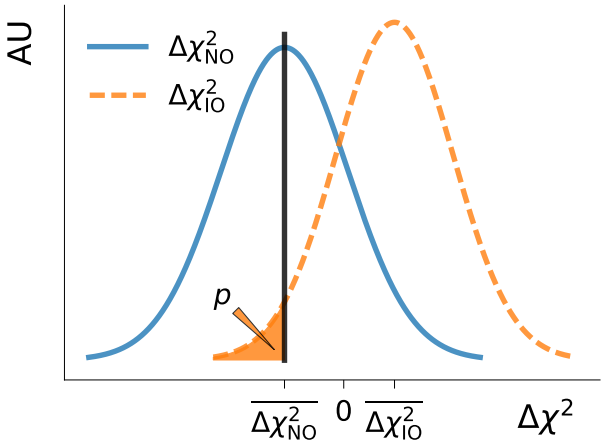


Figure A.12: Example distributions of Equation (A.2). The distribution on the left (solid line) represents the case of NO pseudo-data, while the distribution on the right (dashed) is obtained when the pseudo-data is taken from the IO. Here, $1-p$ corresponds to the confidence level at which the IO is correctly rejected with a probability of 50%.

As the most powerful test statistic for distinguishing two simple hypotheses [32], the *logarithm of the likelihood ratio*

$$\mathcal{T} = -2\ln\left(\frac{\max\limits_{\boldsymbol{\theta}\in\mathrm{NO}} L(\mathbf{n}|\boldsymbol{\mu}(\boldsymbol{\theta}))}{\max\limits_{\boldsymbol{\theta}\in\mathrm{IO}} L(\mathbf{n}|\boldsymbol{\mu}(\boldsymbol{\theta}))}\right) \quad . \tag{A.1}$$

is also useful in assessing the ability of an experiment to discriminate between the two (composite) NMO hypotheses at a given confidence level. It is representative of the degree at which observing the data $\mathbf{n}$ under the NO hypothesis is favored over observing it under the alternate IO hypothesis. The observed spectrum at the detector, $\mathbf{n}$, however, is a convolution of the atmospheric neutrino flux, the effects of neutrino oscillations that bear the NMO signature, the neutrino interaction and detection processes, and the event reconstruction and classification procedure. Each one of these effects is accompanied by systematic uncertainties. As their impact on the predicted spectrum $\boldsymbol{\mu}$ is modeled, the systematic uncertainties directly feed in to the likelihood L of the observation.

For this study, we limit ourselves to a simplified treatment using χ^2 statistics and the *Asimov* dataset. In this approach, the projected median sensitivity is calculated from the average experimental outcomes under the two possible NMO hypotheses, as opposed to performing extensive ensemble tests with randomly fluctuated pseudo-experiments. The log-likelihood expression is a simple χ^2, and Equation (A.1) can be rewritten as the difference

$$\Delta\chi^2 = \chi^2_{\mathrm{NO}} - \chi^2_{\mathrm{IO}} \; . \tag{A.2}$$

Here, χ^2_{NO} is the minimum χ^2 between model predictions and data, with all nuisance parameters profiled out using NO priors (χ^2_{IO} follows analogously).

Parameter	Nominal value		Prior
	NO	IO	
ν_e/ν_μ flux ratio	1.0	1.0	± 0.03
$\nu/\bar{\nu}$ flux ratio	1.0	1.0	± 0.1
Spectral index shift	0.0	0.0	± 0.1
Energy scale	1.0	1.0	± 0.1
Overall normalization	1.0	1.0	± 0.1
θ_{13} (°)	8.5	8.5	± 0.2 [33, 34]
θ_{23} (°)	42.3	49.5	non-Gaussian [33, 34]
Δm^2_{31} (eV^2)	0.00246	-0.00237	$\pm 4.75 \times 10^{-5}$ [33, 34]

Table A.4: Summary of model parameters in the example NMO analysis, including their nominal values for the two NMO hypotheses and Gaussian $\pm 1\sigma$ bounds used as external constraints (priors). The first three parameters are applied to atmospheric neutrino flux predictions from [11], following the procedure laid out in Section 4.2. The values for the three oscillation parameters are based on a recent global fit [33, 34].

An illustration of example distributions of (A.2) for the two different NMO hypotheses is shown in Figure A.12. The goal is to obtain a p-value p which quantifies the statistical compatibility between the hypothesis that is tested and the one assumed to be true. In the ensemble approach, the two distributions would need to be built up by fitting pseudo-experiments. In the Asimov approach, however, certain assumptions about the distribution of (A.2) allow adopting the expression $\sqrt{|\overline{\Delta\chi^2}|}$ as a sensitivity proxy [7], determining the significance at which the wrong ordering can be excluded without the need for pseudo-experiments.

For the profiling of the nuisance parameters (any free model parameters), a numerical algorithm minimizes the χ^2 metric. Whenever external constraints are applied to such parameters, we add those to the χ^2 value as penalty terms (priors). While the presence of these penalty terms is meant to illustrate a typical approach to problems of this sort, their sizes do not follow any precise physical motivation. Table A.4 gives an overview of all used model parameters, their nominal values for NO and IO, and priors (where applied).

Appendix B. Toy Data Model

In the following we provide a parametric toy detector model used to transform the oscillated atmospheric fluxes into event counts. The functions we use either serve as direct inputs (truth) to the various stages of the simulation chain laid out in Section 3, or as sampling distributions from which toy MC samples are drawn. We point out here that these are entirely empirically motivated, and should only be seen as proxies of the performance indicators in VLVνTs (such as the proposed PINGU [1] or KM3NeT/ORCA [3] detectors).

Simplifications or limitations of the model do not affect the computational analysis techniques themselves. Rather, the goal in the following is to capture the most essential features of the expected detector response: threshold effects in detection, the finite accuracy and skew of reconstruction resolution functions, as well as limited flavor and charge identifica-

tion capabilities. This does not invalidate the conclusions drawn from comparing the various analysis approaches.

Appendix B.1. Detection Efficiency

We assume a detector of fiducial mass $M_{\text{fid}} = 10\,\text{megaton}$, with a neutrino detection energy threshold of $E_{\text{th}} = 1\,\text{GeV}$ for all neutrino flavors and interaction channels apart from ν_τ charged current (CC) interactions, where the intrinsic interaction threshold is higher, at $E_{\text{th}} = 3.5\,\text{GeV}$. The detector's effective mass $M^\alpha_{\text{eff}} = \rho_{\text{ice}} V^\alpha_{\text{eff}}$ for a given combination, α, of flavor and interaction type, where ρ_{ice} is the ice density and V^α_{eff} the detector's corresponding effective volume, exhibits a phenomenological dependence on true neutrino energy, E_{true}, asymptotically approaching M_{fid} according to an exponential function:

$$M^\alpha_{\text{eff}}(E_{\text{true}}) = M_{\text{fid}} \times \left(1 - e^{-k_\alpha \times (E_{\text{true}}/\text{GeV} - E_{\text{th}}/\text{GeV})}\right) \text{ for } E_{\text{true}} > E_{\text{th}} \ . \tag{B.1}$$

We include three effective masses to cover all the neutrino interaction channels: one for ν_e, $\bar{\nu}_e$, ν_μ, and $\bar{\nu}_\mu$ CC, one for ν_τ and $\bar{\nu}_\tau$ CC, and one for all NC channels. For the CC channels we choose $k_\alpha = 0.4$, while for the NC channels the function rises more slowly, with $k_\alpha = 0.1$. The left panel of Figure B.13 shows these dependencies for neutrino energies up to $E_{\text{true}} = 80\,\text{GeV}$. The detector can be roughly considered fully efficient ($M_{\text{eff}} = M_{\text{fid}}$) for all detection channels above $E_{\text{true}} \approx 50\,\text{GeV}$.

The ν-$\bar{\nu}$ asymmetry—which is required to make the NMO measurement—will be introduced through differences in flux and cross sections, i.e., it will become apparent in the detector's effective area. The latter we obtain from the effective mass via the conversion

$$A^\alpha_{\text{eff}}(E_{\text{true}}) = \sigma_\alpha(E_{\text{true}}) \times n_{\text{ice}}/\rho_{\text{ice}} \times M^\alpha_{\text{eff}}(E_{\text{true}}) \ , \tag{B.2}$$

where σ_α is the total neutrino-nucleon cross section for a given flavor-interaction channel α, $n_{\text{ice}} \approx 6 \times 10^{23}\,\text{cm}^{-3}$ is the nucleon density in ice, and $\rho_{\text{ice}} \approx 0.92\,\text{g}\,\text{cm}^{-3}$ the mass density.

We also make some simplifying assumptions about the cross sections used in Equation (B.2), in that we take ν_e and ν_μ ($\bar{\nu}_e$ and $\bar{\nu}_\mu$) CC cross sections to be the same, as well as all ν_x ($\bar{\nu}_x$) NC cross sections. In addition, we model all the mentioned cross sections to rise strictly linearly with E_{true} above $E_{\text{true}} = 1\,\text{GeV}$ [35]:

$$\sigma_\alpha(E_{\text{true}})/E_{\text{true}} = c_\alpha \times 10^{-38}\,\text{cm}^2\,\text{GeV}^{-1} \ , \tag{B.3}$$

where we set

$$c_{\nu_{e,\mu}\,\text{CC}} = 2c_{\bar{\nu}_{e,\mu}\,\text{CC}} = 0.70 \ , \tag{B.4}$$

$$c_{\nu_x\,\text{NC}} = 2c_{\bar{\nu}_x\,\text{NC}} = 0.25 \ . \tag{B.5}$$

To obtain ν_τ ($\bar{\nu}_\tau$) CC effective areas, we interpolate the corresponding neutrino-nucleon cross section curves given in [36]. All resulting effective areas as a function of neutrino energy are depicted in the right panel of Figure B.13. We take these to be invariant in azimuth,

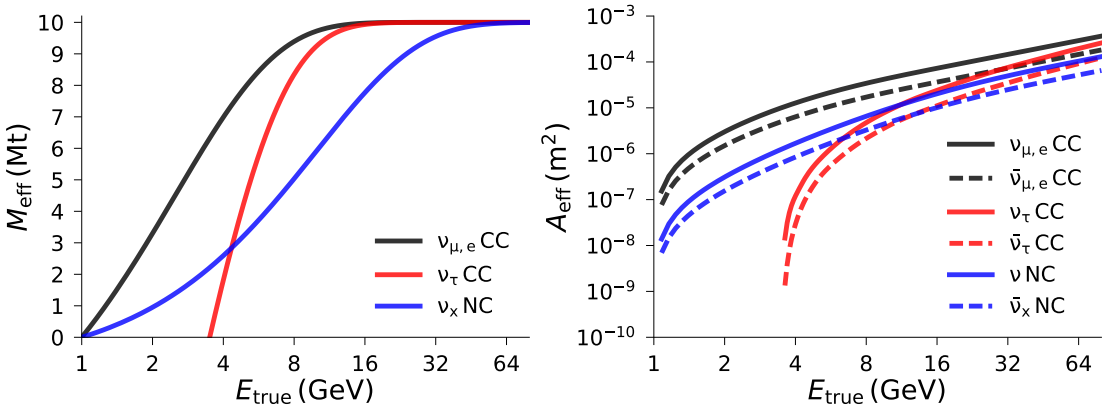


Figure B.13: Effective masses (left) and areas (right) as a function of true neutrino energy for a generic toy detector with fiducial mass of 10 Mt. The dependency of the effective masses on energy is given in Equation (B.1). Cross sections are from Equation (B.3), except for ν_τ and $\bar{\nu}_\tau$ interactions, which are interpolated from [36]. Effective masses are the same for neutrinos and anti-neutrinos. See text for details.

but universally introduce an arbitrary, smooth polynomial modification M with the zenith angle dependency

$$M(x) = \frac{1}{20}(-x^3 + x^2 - x) + 1 \qquad (x \equiv \cos\vartheta_{\text{true}}), \tag{B.6}$$

which we normalize to unit area[13].

Appendix B.2. Reconstruction Resolutions

Neutrino zenith resolutions with respect to $\cos\vartheta$ are represented by single Gaussian distributions. The distributions' parameters are taken as functions of E_{true} only. For each flavor and interaction channel, we assign a mean $\mu_{\Delta\cos\vartheta}(E_{\text{true}}) = 0$ across all energies, and a standard deviation of $\sigma_{\Delta\cos\vartheta}(E_{\text{true}}) = \frac{0.3}{\sqrt{E_{\text{true}}/\text{GeV}}} + 0.05$.

Neutrino energy resolutions we describe using right-skewed Gumbel distributions. These are shifted and scaled by μ' and σ' with respect to their standard form, via the transformation $x \to (x - \mu')/\sigma'$. These parameters again only depend on E_{true}. The CC distributions are assumed identical for all flavors, and are shown in Figure B.14:

$$\mu'^{\text{CC}}_{\Delta E_\nu}(E_{\text{true}}) = 0, \quad \sigma'^{\text{CC}}_{\Delta E_\nu}(E_{\text{true}}) = \left(\frac{0.4}{\sqrt{E_{\text{true}}/\text{GeV}}} + 0.1\right) \times E_{\text{true}} \ . \tag{B.7}$$

For NC interactions, we take a spread that scales with E_{true} in the same way $\sigma'^{\text{CC}}_{\Delta E_\nu}$ does, but assume a non-zero mean due to the undetected energy carried away by the outgoing neutrino: $\mu'^{\text{NC}}_{\Delta E_\nu}(E_{\text{true}}) = -0.6 E_{\text{true}}$.

Note that each energy and cosine zenith residual distribution is renormalized such that its integral over the physical region ($\Delta E_\nu + E_{\text{true}} \geq 0$ and $-1 \leq (\Delta\cos\vartheta + \cos\vartheta_{\text{true}}) \leq 1$) yields 1.

[13] $A_{\text{eff}}(E_{\text{true}})$ is the average over the full sky, $\cos\vartheta_{\text{true}} \in [-1, +1]$.

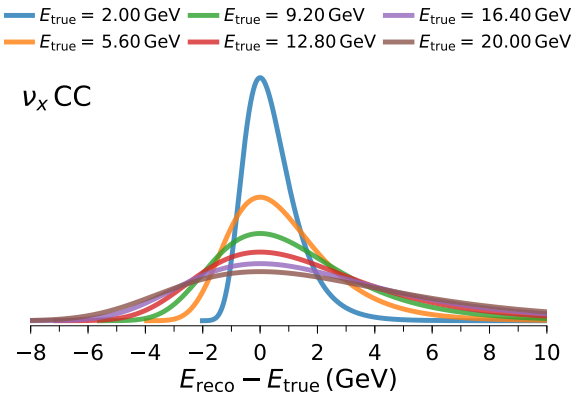


Figure B.14: Example energy resolution functions (right-skewed Gumbel) used for all CC interactions, as given by Equation (B.7). The modes of the corresponding NC resolution functions are shifted by $-0.6E_{\text{true}}$ with respect to the distributions depicted here.

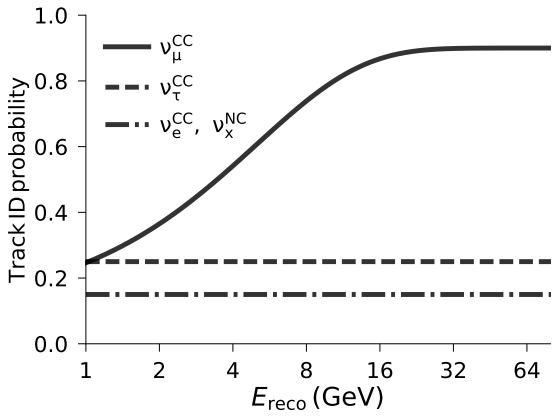


Figure B.15: Event classification efficiencies implemented as functions of reconstructed neutrino energy. Shown is the probability to identify an event of a given type as "track-like". Events are identified as "cascade-like" with complementary probabilities.

Appendix B.3. Event Classification

Correctly identifying few-GeV CC muon neutrino interactions with relatively sparsely instrumented neutrino telescopes in water/ice is difficult mainly for two reasons. The track length of a near minimum ionizing muon is only on the order of a few meters, comparable to the extent of an electromagnetic cascade of the same energy. Also, photon scattering lengths similar to the horizontal spacing between photomultiplier tubes smear out the Cherenkov ring around the muon direction, which is otherwise observed at a specific angle with respect to the muon direction in the medium.

We take into account the muon neutrino CC ("track") identification efficiency $p_{\text{track}}^{\mu,\text{CC}}$ improving with (reconstructed) neutrino energy, E_{reco}, by setting

$$p_{\text{track}}^{\mu,\text{CC}} \equiv p_{\text{track}}^{\mu,\text{CC}}(E_{\text{reco}}) = 0.9 \times \left(1 - e^{-0.2\times(E_{\text{reco}}/\text{GeV}+0.6)}\right) \ . \tag{B.8}$$

This reflects the track length of the secondary muon increasing linearly with its energy, but also the possible production of a low-energy muon which cannot be distinguished from the accompanying hadronic cascade even for higher-energy muon neutrino CC interactions. All other (in)efficiencies are assumed to be constant:

$$p_{\text{track}}^{\text{e,CC}}(E_{\text{reco}}) = p_{\text{track}}^{\text{NC}}(E_{\text{reco}}) = 0.15 \ , \tag{B.9}$$

$$p_{\text{track}}^{\tau,\text{CC}}(E_{\text{reco}}) = 0.25 \ . \tag{B.10}$$

These are shown in Figure B.15. The probability to identify any event as "cascade-like" for a given reconstructed energy is just the complementary probability to that of the track identification.

When a toy MC event is subject to this classification, we assign it one of two discrete numbers—representative of either identification as track or cascade—with the above probabilities.

References

[1] M. G. Aartsen, et al., PINGU: a vision for neutrino and particle physics at the South Pole, J. Phys. G44 (5) (2017) 054006. arXiv:1607.02671, doi:10.1088/1361-6471/44/5/054006.

[2] M. G. Aartsen, et al., Letter of intent: the Precision IceCube Next Generation Upgrade (PINGU), (2017). arXiv:1401.2046v2.

[3] S. Adrian-Martinez, et al., Letter of intent for KM3NeT 2.0, J. Phys. G43 (8) (2016) 084001. arXiv:1601.07459, doi:10.1088/0954-3899/43/8/084001.

[4] R. Barlow, Introduction to statistical issues in particle physics, Statistical problems in particle physics, astrophysics and cosmology. Proceedings, Conference, PHYSTAT 2003, Stanford, USA, September 8-11, 2003, C030908 (2003) MOAT002. arXiv:physics/0311105.

[5] R. H. Byrd, P. Lu, J. Nocedal, A limited memory algorithm for bound constrained optimization, SIAM J. Sci. Comput. 16 (1995) 1190–1208.

[6] G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C71 (2011) 1554, [Erratum: Eur. Phys. J.C73,2501(2013)]. arXiv:1007.1727, doi:10.1140/epjc/s10052-011-1554-0,10.1140/epjc/s10052-013-2501-z.

[7] M. Blennow, P. Coloma, P. Huber, T. Schwetz, Quantifying the sensitivity of oscillation experiments to the neutrino mass ordering, JHEP 2014 (3) (2014) 28. doi:10.1007/JHEP03(2014)028.

[8] K. S. Cranmer, Kernel estimation in high-energy physics, Comput. Phys. Commun. 136 (2001) 198–207. arXiv:hep-ex/0011057, doi:10.1016/S0010-4655(00)00243-5.

[9] D. W. Scott, On optimal and data-based histograms, Biometrika 66 (3) (1979) 605. doi:10.1093/biomet/66.3.605.

[10] F. James, Monte Carlo theory and practice, Rept. Prog. Phys. 43 (1980) 1145. doi:10.1088/0034-4885/43/9/002.

[11] M. Honda, M. Sajjad Athar, T. Kajita, K. Kasahara, S. Midorikawa, Atmospheric neutrino flux calculation using the NRLMSISE-00 atmospheric model, Phys. Rev. D92 (2) (2015) 023004. arXiv:1502.03916, doi:10.1103/PhysRevD.92.023004.

[12] G. D. Barr, T. K. Gaisser, S. Robbins, T. Stanev, Uncertainties in atmospheric neutrino fluxes, Phys. Rev. D74 (2006) 094009. arXiv:astro-ph/0611266, doi:10.1103/PhysRevD.74.094009.

[13] J. Evans, D. G. Gamez, S. D. Porzio, S. Söldner-Rembold, S. Wren, Uncertainties in atmospheric muon-neutrino fluxes arising from cosmic-ray primaries, Phys. Rev. D95 (2) (2017) 023012. arXiv:1612.03219, doi:10.1103/PhysRevD.95.023012.

[14] S. P. Mikheev, A. Y. Smirnov, Resonant amplification of neutrino oscillations in matter and solar neutrino spectroscopy, Nuovo Cim. C9 (1986) 17–26.

[15] L. Wolfenstein, Neutrino oscillations in matter, Phys. Rev. D17 (1978) 2369.

[16] V. Barger, K. Whisnant, S. Pakvasa, R. J. N. Phillips, Matter effects on three-neutrino oscillations, Phys. Rev. D22 (1980) 2718–2726. doi:10.1103/PhysRevD.22.2718.

[17] R. Wendell, Prob3++ software for computing three flavor neutrino oscillation probabilities, http://www.phy.duke.edu/~raw22/public/Prob3++/, 2012.

[18] A. M. Dziewonski, D. L. Anderson, Preliminary reference Earth model, Physics of the Earth and planetary interiors 25 (4) (1981) 297 – 356. doi:http://dx.doi.org/10.1016/0031-9201(81)90046-7.

[19] R. G. Calland, A. C. Kaboth, D. Payne, Accelerated event-by-event neutrino oscillation reweighting with matter effects on a GPU, JINST 9 (2014) P04016. arXiv:1311.7579, doi:10.1088/1748-0221/9/04/P04016.

[20] M. Wallraff, C. Wiebusch, Calculation of oscillation probabilities of atmospheric neutrinos using nuCraft, Comput. Phys. Commun. 197 (2015) 185–189. arXiv:1409.1387, doi:10.1016/j.cpc.2015.07.010.

[21] T. K. Gaisser, T. Stanev, Path length distributions of atmospheric neutrinos, Phys. Rev. D57 (1998) 1977–1982. doi:10.1103/PhysRevD.57.1977.

[22] I. S. Abramson, On bandwidth variation in kernel estimates—a square root law, Ann. Statist. 10 (4) (1982) 1217–1223. doi:10.1214/aos/1176345986.

[23] Z. I. Botev, J. F. Grotowski, D. P. Kroese, Kernel density estimation via diffusion, Ann. Statist. 38 (5) (2010) 2916–2957. doi:10.1214/10-AOS799.
[24] S. J. Sheather, M. C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, Journal of the Royal Statistical Society. Series B (Methodological) 53 (3) (1991) 683–690.
[25] Y. Fukuda, et al., Evidence for oscillation of atmospheric neutrinos, Phys. Rev. Lett. 81 (1998) 1562–1567. arXiv:hep-ex/9807003, doi:10.1103/PhysRevLett.81.1562.
[26] Q. R. Ahmad, et al., Measurement of the rate of $\nu_e + d \rightarrow p + p + e^-$ interactions produced by ^8B solar neutrinos at the Sudbury Neutrino Observatory, Phys. Rev. Lett. 87 (2001) 071301. arXiv:nucl-ex/0106015, doi:10.1103/PhysRevLett.87.071301.
[27] B. Aharmim, et al., Combined analysis of all three phases of solar neutrino data from the Sudbury Neutrino Observatory, Phys. Rev. C88 (2013) 025501. arXiv:1109.0763, doi:10.1103/PhysRevC.88.025501.
[28] C. Patrignani, et al., Review of particle physics, Chin. Phys. C40 (10) (2016) 100001. doi:10.1088/1674-1137/40/10/100001.
[29] S. T. Petcov, S. Toshev, Three neutrino oscillations in matter: analytical results in the adiabatic approximation, Phys. Lett. B187 (1987) 120–126. doi:10.1016/0370-2693(87)90083-9.
[30] E. K. Akhmedov, M. Maltoni, A. Yu. Smirnov, Oscillations of high energy neutrinos in matter: precise formalism and parametric resonance, Phys. Rev. Lett. 95 (2005) 211801. arXiv:hep-ph/0506064, doi:10.1103/PhysRevLett.95.211801.
[31] K. Abe, et al., Letter of intent: the Hyper-Kamiokande experiment — detector design and physics potential —. arXiv:1109.3262.
[32] J. Neyman, E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 231 (694-706) (1933) 289–337. doi:10.1098/rsta.1933.0009.
[33] M. C. Gonzalez-Garcia, M. Maltoni, T. Schwetz, Updated fit to three neutrino mixing: status of leptonic CP violation, JHEP 2014 (11) (2014) 52. doi:10.1007/JHEP11(2014)052.
[34] NuFIT 2.0 (2014), www.nu-fit.org.
[35] J. A. Formaggio, G. P. Zeller, From eV to EeV: neutrino cross sections across energy scales, Rev. Mod. Phys. 84 (2012) 1307–1341. arXiv:1305.7513, doi:10.1103/RevModPhys.84.1307.
[36] A. Gazizov, M. Kowalski, K. S. Kuzmin, V. A. Naumov, C. Spiering, Neutrino-nucleon cross sections at energies of megaton-scale detectors, EPJ Web Conf. 116 (2016) 08003. arXiv:1604.02092, doi:10.1051/epjconf/201611608003.